UNIVERSITY OF CALIFORNIA

Santa Barbara

Understanding and Mitigating Search Errors in 3D Volumetric Images

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychological and Brain Sciences

by

Devi S. Klein

Committee in charge:

Professor Miguel Eckstein, Chair

Professor Craig Abbey

Professor Barry Giesbrecht

Professor Thomas Sprague

June 2024

The dissertation of Devi S. Klein is approved.

_____

Craig Abbey

_____

Barry Giesbrecht

_____

Thomas Sprague

_____

Miguel Eckstein, Committee Chair

June 2024

Understanding and Mitigating Search Errors in 3D Volumetric Images

by

Devi S. Klein

ACKNOWLEDGEMENTS

To say pursuing a Ph.D. is a marathon, not a sprint, encapsulates my journey through graduate school. Obstacles will be there along the way, and perseverance and tenacity are requisite to push through the final few miles to the finish line. But along the way, having the supporting cast that is there cheering you on, guiding you, and diminishing your self-doubt is the true x-factor in crossing the finish line. Reflecting on my time at the University of California, Santa Barbara, I am forever indebted to my lab members (past and present), advisors, committee members, and friends I have made along the way. This diverse group of people has not just supported me, but they have been the wind beneath my wings, keeping me motivated, and continually pushing me both directly and indirectly to continue to learn and contribute to the field of medical image perception. The body of work presented herein is an amalgamation of ideas formulated from conversations, philosophical debates, and strong mentorship from my advisor, Dr. Miguel P. Eckstein.

So, naturally, to begin my acknowledgments, I'd like to express my deepest gratitude to my advisor, Miguel Eckstein. His unwavering belief in my potential, his continual push for me to take on new projects, his guidance in refining my analytical and critical thinking skills, and his effort in producing exceptional research have not only shaped my academic career but also transformed me as a person. Having an advisor who does not accept the bare minimum but only solid, rigorous empirical work has led me to appreciate what it really means to push the field forward. The countless hours spent with him generating experiment ideas, controlling for confounding factors, and pursuing a comprehensive list of data

analysis checks to ensure the validity of our results have cultivated in me a deep appreciation for how to become a diligent scientist.

Next, I would like to thank Dr. Craig Abbey. He played an integral role in teaching me about mathematical model observers and task-based medical image quality assessment. Without his mentorship and countless hours helping me revise Chapter III of this thesis, the manuscript would not have been as polished or mathematically correct. Additionally, he has taught me what it means to write a succinct and effective paper Introduction. Specifically, he proposed that one needs to answer the following three questions: (1) what is your research question? (2) what is your approach? and (3) Why is this question important, and why is this a good approach to answer it? Everything else in an introduction is superfluous because you want to get the reader from point A to point B (the methods sections) as quickly as possible.

Lastly, in academia, I would like to thank all my previous and current lab mates. You guys have been incredible in shaping my critical thinking skills and learning new analysis techniques, and have ultimately brought joy to me along this arduous journey. You are all lifelong friends, peers, and mentors who deserve the best in your future endeavors. Beyond the lab, I would like to thank all the incredible graduate students I was fortunate enough to befriend. There are too many fond memories and good times to enumerate here, but you have all shown me the importance of taking time away from work to acquire a healthy "work-life balance" that is critical for succeeding in graduate school. I am very thankful for your support, the laughs, the tears, and the adventures!

Finally, I would like to express my heartfelt gratitude to my family. First, my parents have served as my rock throughout this journey. Since day 1, they have provided me with unwavering support. I cannot overstate how important they have been to me throughout

graduate school. My two brothers and my grandparents were also there, always cheering me on and providing unconditional support. Lastly, my fiancé, Cate, you have been such an incredible person. You have been patient and supportive, and your belief in me to succeed has been such a huge contributing factor to my motivation. I cannot wait for our future together. Your love and support have been the foundation of my strength and resilience during this Ph.D. journey.

VITA OF Devi S. Klein
June 2024

EDUCATION

Bachelor of Science in Psychology, University of Washington, December 2016
Doctor of Philosophy in Psychological and Brain Sciences, University of California, Santa Barbara, June 2024 (expected)

PROFESSIONAL EMPLOYMENT

2017-2024: Graduate Student Researcher/Teaching Assistant, Department of Psychological and Brain Sciences, University of California, Santa Barbara
Summer 2022: Summer Internship, Nvidia

PUBLICATIONS

D. S. Klein, M. A. Lago, C. K. Abbey, and M. P. Eckstein, "A 2D Synthesized Image Improves the 3D Search for Foveated Visual Systems," in *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2176-2188, Feb. 2023, doi: 10.1109/TMI.2023.3246005.

Han, N. X., Srivastava, S., Xu, A., **Klein, D.**, & Beyeler, M. "Deep Learning-Based Scene Simplification for Bionic Vision," Augmented Humans International Conference, Feb. 2021 doi: https://doi.org/10.1145/3458709.3458982.

Devi S. Klein, Miguel A. Lago, Miguel P. Eckstein, "The perceptual influence of 2D synthesized images on 3D search," Proc. SPIE 11599, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment, 115990P Feb. 2021, doi: https://doi.org/10.1117/12.2582262.

MANUSCRIPTS UNDER REVIEW

Klein, D., Eckstein, M. P., "The Search Termination Criterion Mediating Under Exploration of 3D Volumetric Images." PsyArXiv, Aug. 2023, doi: https://doi.org/10.31234/osf.io/duyw8.

Klein, D., Karmakar, S., Jonnalagadda, A., Abbey, C. K., Eckstein M. P, "Greater benefits of deep learning-based computer-aided detection systems for finding small signals in 3D volumetric medical images." ArXiv, Apr. 2024, doi: https://doi.org/10.48550/arXiv.2405.00144

MANUSCRIPTS IN PREPARATION

Klein, D., Eckstein, M. P., "More than meets the (single) eye: the greater benefits of group decision-making for visual search in large 3D volumetric medical images."

Klein, D., Spjut, J. Boudaoud, B., Kim, J., "The Influence of Variable Frame Timing on First-Person Gaming, Jun. 2023, https://doi.org/10.48550/arXiv.2306.01691.

CONFERENCE

POSTERS

Devi Klein, Miguel P. Eckstein; Sufficient eye movement coverage of the 2D image plane might mediate under-exploration in 3D search. *Journal of Vision* 2023; 23(9):5831, doi: https://doi.org/10.1167/jov.23.9.5831.

Devi Klein, Miguel P. Eckstein; Meta-awareness of anisotropic processing of visual information across the visual field. Journal of Vision 2022; 22(14):4347, doi: https://doi/org/10.1167/jov.22.14.4347.

Devi S. Klein, Miguel A. Lago, Miguel P. Eckstein; Guiding search in 3D volumes with 2D synthesized images. *Journal of Vision* 2021; 21(9):2919, doi: https://doi.org/10.1167/jov.21.9.2919.

Aiwen Xu, Nicole Han, Sudhanshu Srivastava, **Devi Klein**, Michael Beyeler; Enhancing simulated prosthetic vision with deep learning-based scene simplification strategies. *Journal of Vision 2021*; 21(9):2308, doi: https://doi.org/10.1167/jov.21.9.2308.

TALKS

Devi S. Klein, Miguel A. Lago, Miguel P. Eckstein, "The perceptual influence of 2D synthesized images on 3D search," Proc. SPIE 11599, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment, 115990P Feb. 2021, doi: https://doi.org/10.1117/12.2582262.

Han, N. X., Srivastava, S., Xu, A., **Klein, D.**, & Beyeler, M. "Deep Learning-Based Scene Simplification for Bionic Vision," Augmented Humans International Conference, Feb. 2021 doi: https://doi.org/10.1145/3458709.3458982.

FIELDS OF STUDY

Major Field: Human Visual Perception

Studies in Visual Search with Professor Miguel Eckstein

Studies in Computational Modeling, Medical Image Perception, and Observer Performance with Professor Miguel Eckstein and Dr. Craig Abbey

Studies in Bionic Vision with Professor Michael Beyeler

ABSTRACT

Understanding and Mitigating Search Errors in 3D Volumetric Images

by

Devi S. Klein

In the field of oncology, three-dimensional volumetric medical images provide radiologists with a detailed visual representation of various anatomical structures that facilitate the early detection and characterization of malignant lesions but at the cost of an increased search space. Recent work (Lago, Jonnalagadda, et al., 2021) establishes that human observers rely heavily on peripheral visual processing away from the point of fixation when searching for signals in 3D volumetric images. The searcher's over-reliance on peripheral vision interacts strongly with how much of the volume they explore and with how much they report they have explored. Specifically, observers under-explore—as determined by the percentage of the volume covered by the Useful Field of View (UFOV)—and overestimate the percentage of volume they explored through self-report measures. Consequently, they miss small signals during the search. This thesis aims to elucidate the psychological factors mediating human under-exploration of 3D volumetric image data.

The second thrust of this thesis is to investigate three solutions to mitigate the detrimental impact of under-exploration in 3D images. The first method is a 2D synthetic view of the 3D data that observers can utilize as additional information when performing the 3D search. I establish through behavioral measurements and a computational model

simulating foveated vision how the 2D-S guides eye movements to suspicious regions in the 3D volume. In turn, this guidance allows observers to find the small signal that would otherwise be missed without the 2D-S adjunct. The second method involves a different type of search aid, a convolutional neural network, which acts as a computer-aided detection system to assist human observers during the 3D search. Like the 2D-S, it guides eye movements to suspicious regions in a 3D volumetric image that observers would have otherwise not looked at.

The last method is inspired by the power of group decision-making. It investigates how combining multiple independent judgements from a group of searchers can lead to more exploration of the search space and a higher chance of detecting the small signal. Together, the body of work herein provides empirical results from laboratory studies to further our understanding of how humans interact with 3D imaging modalities with the goal of improving healthcare services relating to early cancer screenings.

TABLE OF CONTENTS

I. Introduction

1.1. The perception of medical images

Since Wilhelm Rontgen took the first X-ray photograph of his wife's hand in 1895, medical imaging technology has made remarkable progress in imaging the human body. From plain film radiography to Computed Tomography (Bercovich & Javitt, 2018), the advancements have been significant (L. Zhou et al., 2022). Medical imaging has not only improved surgery with interventional imaging techniques such as coronary angiography (La Vecchia, 2013), but it has also revolutionized the early detection of diseases like cancer (Pisano et al., 2005; Sharma et al., 2012) and diabetic retinopathy (Salz & Witkin, 2015). Medical image has been the cornerstone of cognitive neuroscience since the advent of functional magnetic resonance imaging in the early 90s (Ogawa et al., 1990). The impact of medical imaging on modern-day society cannot be overstated.

However, any acknowledgment of medical imaging advancements requires a discussion of image quality, as it is part and parcel of assessing a medical imaging device's functional use (i.e., its ability to provide information and reduce diagnostic uncertainty). For instance, in radiology, a great deal of research is spent on mapping the relationship between ionizing radiation dosage and image quality, the latter of which can be characterized by spatial resolution, contrast, and photon noise (Huda et al., 2002). Image reconstruction algorithms can further enhance the image by increasing diagnostic information (Gothwal et al., 2022), and there are two schools of thought for measuring medical image quality. The first school of thought evaluates image quality based on fidelity metrics such as Peak Signal-to-Noise Ratio and Mean Squared Error (Chow & Paramesran, 2016; Samajdar & Quraishi, 2015).

The second school of thought is rooted in task-based image quality assessment, and it provides a more objective avenue for judging the functional use of an imaging system (Barrett, 1990; Barrett & Myers, 2013). This approach requires defining the task (e.g., estimating some property of the image or classifying an image as signal or noise), an observer (a human or mathematical model), and a figure of merit (e.g., sensitivity, specificity, or the area under the receiver operating curve). This thesis follows the latter convention for assessing image quality, focusing on classification tasks with human observers and evaluating their performance using multiple figures of merit.

Human observer performance studies remain the gold standard for assessing the quality of medical images because clinicians make the final interpretation and recommendation for the next steps. However, clinicians are not always perfect; they are prone to making perceptual and decision errors (Krupinski, 2010, 2011). There is, in fact, a great deal of variability in radiologist performance (Beam et al., 2003), and a substantial amount of research has been conducted to understand the role of expertise in mediating diagnostic performance (Nodine & Mello-Thoms, 2010; Waite et al., 2019). However, to assess the functionality of an imaging system, it is essential to not only understand the role of expertise but also to understand how humans perform visual searches in medical images and taxonomize the types of visual-cognitive errors they make while interpreting them.

One common error arises from the interaction between search and prevalence-low target prevalence leads to misses (false negatives). This finding has been attributed to early termination of the search rather than perceptual or identification errors, suggesting that priors or expectations about target prevalence guide the search to some degree before the visual interpretation of the image begins (Fleck & Mitroff, 2007; Mitroff & Biggs, 2014;

Wolfe et al., 2005). Satisfaction of search is a related cognitive error whereby after detection of the first anomalous feature, a searcher becomes satisfied and has a higher likelihood of missing other signs of malignancy in the image (Adamo et al., 2021; Berbaum et al., 1990; Tuddenham, 1962). Other notable factors related to radiologist workflow include fatigue, time of day, and eye strain, all of which can negatively impact performance (Krupinski et al., 2010; Taylor-Phillips & Stinton, 2019b).

Visual-cognitive errors directly related to the perception of an image have been identified with eye-tracking and characterized from a vision science perspective (Krupinski, 1996; Kundel et al., 1978). Kundel, Nodine, and Carmody provided one of the first studies categorizing misses of malignant lesions based on where a radiologist looked in an image. They labeled misses as recognition errors, decision errors, or search errors. Recognition errors were defined as short fixations on the lesion and target-absent responses, and decision errors required a longer fixation dwell time on the lesion (> 500ms). These two types of errors have been attributed to visual masking effects and human observers' inability to see through the noise and backgrounds in medical images (Burgess et al., 1997, 2001; Mello-Thoms et al., 2005). Search errors occurred when the radiologist failed to fixate the signal and reported it as absent. Understanding the nature of search errors, why they arise, and assessing methods for mitigating them is of central interest to this thesis.

Search errors occur because of the interaction between eye movement exploration during search and the foveated nature of the human visual system. Visual information presented at the fovea or near the gaze point is processed with high resolution (Levi et al., 1985; Robson & Graham, 1981; Rovamo et al., 1984) because of a densely packed array of cone photoreceptors in the foveola (Curcio et al., 1990), one-to-one mapping between cones to

retinal ganglion cells via bipolar cells (Curcio & Allen, 1990), and the disproportionate amount of neurons per mm^2 in the primary visual cortex dedicated to processing foveal information (Duncan & Boynton, 2003). Peripheral vision, on the other hand, which processes most of the information in the visual field, is characterized by lower spatial resolution but still plays a significant role in guiding eye movements (Rosenholtz, 2016; Stewart et al., 2020). Small signals, comprised of high spatial frequency information, which are hard to detect in the visual periphery (Lago, Sechopoulos, et al., 2020), are the most prone to becoming a search error if a radiologist fails to visually scrutinize a medical image sufficiently.

1.2. From 2D to 3D medical images

To this point, the discussion has broadly been concerned with medical images and how radiologists interpret them. No distinction has been made regarding the interpretation of 2D versus three-dimensional medical images. In the field of oncology, three-dimensional volumetric medical images are becoming the standard for interpretation because they reduce tissue superposition inherent in 2D projection images. The reduction in tissue overlap diminishes partial occlusion of signals of interest by surrounding dense tissue, thus providing radiologists with a detailed visual representation of various anatomical structures that facilitate the early detection and characterization of malignant lesions (Alabousi et al., 2020; Gould, 2014).

As an example, for early breast cancer detection, screenings have evolved over the past decade from the interpretation of 2D mammograms to the interpretation of digital breast tomosynthesis (DBT) volumetric data (Georgian-Smith et al., 2019; Skaane, 2017). The use

of DBT has been shown to improve early cancer detection (Badano et al., 2018; Georgian-Smith et al., 2019).

DBT generates a stack of cross-sectional "slices" of the breast from reconstructed X-ray projections acquired over a limited arc range (i.e., quasi-3D view) as a detector and X-ray source revolve around the patient (Sechopoulos, 2013). The slices are viewed one at a time as part of a sequence of images displayed on a computer monitor, permitting radiologists to scroll back and forth through the third dimension of the volume. Displaying the image data in this way allows radiologists to better segment abnormal tissue from the surrounding parenchyma that may otherwise be partially or fully occluded if the patient's anatomy were to be interpreted from a two-dimensional projection image (e.g., full-field digital mammogram) (Helvie, 2010).

The benefits associated with the additional depth information in 3D volumetric images come with an unprecedented increase in the search space (50-90 slices per scan (Baker & Lo, 2011; Gur et al., 2009)), which can increase signal position uncertainty. Depending on the 3D image reconstruction algorithm and parameter settings (e.g., z resolution and slice thickness), small signals of interest can appear on only one or a few of the total slices in the 3D stack (Williams & Drew, 2019). The large search space can also affect the reader. It would be prohibitively time-consuming to scan exhaustively, with eye movements, each cross-sectional slice in the stack of images before terminating one's search. In fact, given average reading times of DBT images that range from 2-3 minutes per scan (Good et al., 2008; Gur et al., 2009) and typical fixation durations of 250-350 ms, it is estimated that radiologists would max out at approximately 14 fixations per slice, which may negatively impact their ability to find signals of interest.

How to optimize the interpretation of 3D images to increase performance further is still an active area of research (Drew, Vo, Olwal, et al., 2013; Drew, Vo, & Wolfe, 2013; Lago, Jonnalagadda, et al., 2021; Rubin et al., 2015; Williams & Drew, 2019). Eye-tracking studies have revealed how radiologists search through 3D images. Some radiologists scan a cross-sectional slice before moving on to the next slice, while others tend to fixate on one location and drill through the 3D stack of images before refixation somewhere else (Drew, Vo, Olwal, et al., 2013). Eye-tracking studies have also shown how a large search space can negatively affect perceptual performance vis-à-vis under-exploration of the 3D volume, whereby observers do not exhaustively direct their center of gaze to every region in the space.

For example, recent work (Lago, Jonnalagadda, et al., 2021) establishes that human observers rely heavily on peripheral visual processing away from the point of fixation when searching for signals in 3D volumetric images. The searcher's over-reliance on peripheral vision interacts strongly with how much of the volume they explore and how much they report they have explored. Specifically, observers under-explore—as determined by the percentage of the volume covered by the Useful Field of View (UFOV)—and overestimate the percentage of volume they explored through self-report measures.

Under-exploration in 3D images differentially impacts the detection of signals based on their spatial size. Large signals more detectable in the visual periphery are less affected by under-exploration. On the other hand, due to the foveated nature of the human visual system (i.e., low spatial acuity in the visual periphery), observers miss small signals (search errors) embedded in the 3D space. These small signals are otherwise salient and easily detected in a

2D image because observers can direct their fovea to most regions in the image in a time-efficient manner.

The study by Lago et al. identified a unique visual-cognitive error related to 3D imaging modalities. However, it is not well understood why observers under-explore during 3D search and overestimate how much they explore. Additionally, it remains unclear what the best solutions are to mitigate 3D search errors for small signals. Various techniques have been adopted in clinical practice to complement radiologist decisions and mitigate errors. One simple solution is to add a second reader (Ciatto et al., 2005; Duijm et al., 2004). In many countries, radiological images undergo double reading (Taylor-Phillips & Stinton, 2019a), and there are theoretical accounts of how pooling multiple radiologist's decisions can outperform the average radiologist (Brennan et al., 2019) and even the best-performing member (Kurvers et al., 2016; Wolf et al., 2015).

In the US, independent double reading in mammography is seldom done. Nonetheless, for many years, mammography has used computer-aided detection (CADe) and classification (CADx) to work as a "second reader" to identify potential malignancies (Doi, 2007; Giger et al., 2008). More recently, artificial intelligence-based CADx can be incorporated into the workflow at the radiologist's discretion, such as by filtering out exams with the lowest likelihood of malignancy (triage), thus saving time and effort for more ambiguous images (Rodriguez-Ruiz et al., 2019). In general, the CADe prompts are superimposed on the medical image, and readers can visually attend to those locations while interpreting the image.

Finally, since the advent and FDA approval of DBT in 2011, radiologists have either a mammogram or a 2D synthetic view (2D-S) of the corresponding DBT data available to aid

their visual search in the 3D volumetric image. More experienced radiologists have developed visual search strategies that are optimized for interpreting 2D images (Wolfe et al., 2016), and there are documented cases of them being able to glean relevant information quickly from a 2D medical image (Drew, Evans, et al., 2013; Kundel & Nodine, 1975). Thus, the 2D image should reduce or mitigate errors that may arise by just reading a 3D image alone.

Although studies have shown how CAD and double reading influence screening mammography with 2D images, less has been done to evaluate their influences on 3D images and search errors. In the case of complementary 2D synthetic images accompanying 3D images, there is no systematic vision science investigation explicating how it improves 3D search or reduces search errors.

In summary, this thesis will increase our understanding of why humans under-explore 3D volumetric images and provide potential solutions to mitigate 3D search errors with applications to radiology. To achieve this, the studies outlined below will focus on addressing the following questions. When observers under-explore 3D images, what evidence do they base their quitting decision on? How do the 2D synthetic image and CADe influence observers' search strategies in 3D? Are there unique benefits gained from aggregating observers' judgments in 3D that are not seen in 2D searches? For each of the three methods (2D synthetic view, CADe, or group decisions), which types of signals benefit the most? Lastly, which of these techniques is best suited to mitigate 3D search errors?

1.3. Organization of this thesis

This thesis aims to better understand why humans under-explore 3D volumes and investigate solutions to mitigate the detrimental impact of this search strategy. Chapter II elucidates possible visual-cognitive mechanisms that subserve the under-exploration of 3D volumes. Specifically, I utilize psychophysical measurements and eye-tracking to derive estimates of the proportion of area explored with the perceived UFOV (how well observers think they can see in the visual periphery) versus the empirical UFOV (how well they can see in the visual periphery) to contrast two plausible search-termination thresholds human observers utilize after reporting the absence of the signal.

The remainder of this thesis investigates three ways of reducing the detrimental impact of an overreliance on peripheral vision while performing the 3D search. Chapter III assesses the impact of a 2D-S, a 2D synthesized view of the 3D volume, serving as an adjunct to the 3D search. I establish through behavioral measurements and a computational model simulating foveated vision how the 2D-S guides eye movements to suspicious regions in the 3D volume. In turn, this guidance allows observers to find the small signal that would otherwise be missed without the 2D-S adjunct. Chapter IV homes in on a second type of search aid, a convolutional neural network, which acts as a computer-aided detection system to assist human observers during the 3D search. Like the 2D-S, it guides eye movements to suspicious regions in a 3D volumetric image. Chapter V employs a modified wisdom of the crowd model, a majority vote with exception rule. We establish that many observers visually scanning the same 3D volume outperform any single observer in the group if at least one person fixates on the signal of interest and the exception rule supplants the group's decision with their explicit judgment.

II. Relating the perceived useful field of view to visual search with 2D images and 3D volumetric images

2.1. Abstract

3D volumetric images are prevalent in industries ranging from radiology and oncology to airport luggage screenings. With 3D medical images, however, radiologists are visually overburdened by the vast amount of image data requiring inspection. Radiologists and trained (non-radiologist) observers often under-explore 3D images with eye movements—producing misses of small targets undetectable in the visual periphery. We investigate why observers under-explore 3D images by quantifying their perceived exploration of the search area and relating it to the extent of their search on trials where they report "target-absent." Six trained observers participated in two eye-tracking experiments to evaluate whether the area explored by the Useful Field of View (UFOV) influences the under-exploration of the 3D images. Experiment 1 estimated empirical and perceived target-specific UFOVs per observer. We tested the observer's detectability of a small and large target embedded in 1/f noise as a function of retinal eccentricity (empirical UFOV). Observers also estimated how well they could see the two targets in their visual periphery (perceived UFOV). Experiment 2 had observers participate in a 2D and a 3D search for the two targets. The area explored with the perceived but not empirical UFOVs was consistent across targets in the 2D and 3D searches. Moreover, while performing the 3D search, people covered the 2D image plane with their perceived UFOVs to the same extent as in the 2D search task, leaving much of the 3D volume unexplored.

2.2. Introduction

Three-dimensional (3D) volumetric imaging technology is becoming a mainstay in industries ranging from medical imaging (Williams & Drew, 2019) to airport screening (Parker et al., 2022). 3D volumetric imaging diminishes occlusion and object superposition inherent in traditional 2D imaging systems because a rotating energy source (e.g., X-ray tube) interacts with the object of interest at various viewing angles. In computed tomography, for example, a set of sinograms generated from an X-ray source positioned at different angles to the human body can be integrated via filtered back projection to produce a 3D view of the lungs, liver, or other anatomical structures (Schofield et al., 2020, p. 1). Similar 3D reconstruction algorithms are used for early breast cancer detection. For example, digital breast tomosynthesis (DBT or 3D mammography) minimizes tissue superposition (Sechopoulos, 2013) and can increase sensitivity and specificity relative to 2D full-field digital mammography (Skaane, 2017).

Regardless of the 3D imaging technology or anatomy under view, visual search in medical images is a difficult task that has garnered much interest (Krupinski, 2000). Additional complexity is introduced when considering how to display the 3D data to a human observer (Calhoun et al., 1999; Getty & Green, 2007; Lu & Sakamoto, 2018; Maupu et al., 2005; Rubin et al., 1996). It is common practice to render the 3D volumetric image as a stack of 2D image "slices." Each slice represents a cross-sectional view of the anatomy. Therefore, the reader must scroll back and forth through the third dimension of the volume at their own pace and visually inspect each slice as they appear one at a time on the computer monitor. One might ask if there are drawbacks to this 3D presentation strategy. Specifically, do visual-cognitive bottlenecks hinder search performance in the 3D image?

Recent work has investigated this question directly by testing how human observers search for spatially large and small targets in both 3D volumetric images and 2D images (Lago, Abbey, et al., 2021a; Lago, Jonnalagadda, et al., 2021). Lago et al. identified that the small target was readily detected in 2D images because (1) it was salient when fixated and (2) humans can direct their fovea to most regions of the image in a time-efficient manner. Whereas in the 3D search, the small target was often missed. They hypothesized that the low detectability of the small target in the visual periphery and the non-exhaustive coverage of the image data with the eye movements caused observers to miss the small target. They corroborated this hypothesis by examining the 3D scan paths of observers while they performed the search and found that a large fraction of misses occurred because observers failed to foveate the target. Furthermore, requiring observers to extend their 3D search exploration time caused a reduction in errors.

For the larger target, which was more detectable in the visual periphery, they found no performance detriment for 3D images due to search errors. This was in spite of the fact that observers' explorations were shorter, and their eye movements covered less of the volumetric area compared to the 3D search of the small target. Together, this observed interaction between search modality and target type further supported their hypothesis that related the 3D search errors to the interaction between under-exploration and the peripheral detectability of the target. This 3D under-exploratory behavior can have clinical significance because the results were replicated in the same study with radiologists and 3D DBT phantoms (Lago, Jonnalagadda, et al., 2021), and similar results of under-exploration were reported in a study of radiologist searching through lung CT scans (Rubin et al., 2015).

The outstanding question remains—why do observers under-explore 3D volumetric images relative to 2D images? Relatedly, why do observers change their search exploration for different types of targets? Many factors influence an observer's stopping criterion during the 2D search. Finding an easily detectable target produces a self-terminating search rather than an exhaustive scan of all items in the visual display (Van Zandt & Townsend, 1993). Lower prevalence can decrease the search times (Ishibashi et al., 2012), while high rewards can also increase them (but see (Wolfe, 2012) for a discussion of the evidence of no effects of rewards). Conceptual models of search posit a quitting threshold based on an estimate of the "effective set size" or number of candidate target items searched (Wolfe, 2012) and an estimate of the ease with which candidate targets can be localized and discounted during covert deployment of attention (Chun & Wolfe, 1996; Wolfe, 2021). Models of eye movements during 2D search implicitly incorporate stopping criteria based on the accumulated evidence of the presence of the target (Akbas & Eckstein, 2017; Lago, Abbey, et al., 2021a; Najemnik & Geisler, 2005) or a stopping criterion based on the proportion of the image area explored (Lago, Abbey, et al., 2021a). Thus, one possible explanation for the under-exploration of 3D volumetric images and why exploration differs across targets is that each imaging modality (2D vs. 3D) and target type has a different search-termination threshold that involves multiple factors.

A different and more parsimonious possibility is that observers use a common metric across target types and 2D and 3D image modalities to terminate their search. This paper investigates whether an observer's internal estimate of the proportion of area explored, sans finding the target, serves as that common metric. To quantify the proportion of the area explored, one can use the Useful Field of View (UFOV) construct (Ball et al., 1988;

Hulleman & Olivers, 2017; Lago, Sechopoulos, et al., 2020; Wu & Wolfe, 2019) and observers fixation patterns during their search. The UFOV is defined by the area around the fovea for which a target is detected with a high probability. For example, following classic conventions (Drew, Vo, Olwal, et al., 2013; Kundel et al., 1989), Lago et al. chose a single UFOV for both types of targets, a circular area with a radius of 2.5° visual angle centered at each recorded fixation position during the search. They found that humans, on average, explored a different proportion of the area when tasked to look for the two targets and across 2D/3D images (Lago, Jonnalagadda, et al., 2021). Thus, their data does not support the idea that an observer compared the proportion of the area explored to a single stopping criterion to gauge when to terminate their search.

However, one limitation of their approach is the assumption of a theoretical UFOV of 2.5°. Studies have demonstrated that empirically measured UFOVs differ across targets and backgrounds (Carmody et al., 1980; Ebner et al., 2017; Lago, Sechopoulos, et al., 2020). A target-specific empirical UFOV measurement quantifies the net effect of factors influencing foveal and peripheral detectability (Banks et al., 1991), such as the interplay between target and background spatial frequencies (Abbey & Eckstein, 2007; Burgess et al., 2001; Lago, Abbey, et al., 2021b), masking and crowding (Bouma, 1970; Pelli et al., 2004; Rosenholtz, 2016; Strasburger et al., 2011; Vater et al., 2022). Thus, if we incorporate target-specific empirical UFOVs (rather than the generic 2.5° theoretical UFOV) with the measured search fixations, observers might be exploring a similar proportion of area when tasked to look for each target, which would suggest a common mediating the stopping criterion.

The hypothesis that observers might use the proportion of area explored with their UFOV can theoretically relate to other conceptual models that base the stopping criterion on

the number of candidate target items examined (Van Zandt & Townsend, 1993; Wolfe, 2012). However, rather than the number of candidate items examined, the proportion of image/volume area explored with the UFOV quantifies the image's area examined. The measure is particularly appropriate for textures (Abbey & Eckstein, 2014; Burgess et al., 2001; Castella et al., 2008) or medical images (Bochud et al., 2004) such as Gaussian noise images and mammograms/DBT images, and scenes (Neider & Zelinsky, 2008; Akbas & Eckstein, 2017), for which the number of items is difficult to define (Hulleman & Olivers, 2017).

Two crucial considerations factor into an observer's ability to compute the proportion of the area explored with their target-specific empirical UFOV. First, observers need to keep track of which areas they have already explored during their search. Despite studies demonstrating that observers do not have a good memory of the exact locations they previously fixated/searched (Horowitz & Wolfe, 1998; Võ et al., 2016), although this has been debated (Beck et al., 2006; Peterson et al., 2001), there is evidence that they can provide estimates that approximate the proportion of area explored in 2D search tasks (Lago, Jonnalagadda, et al., 2021). Specifically, when observers were probed at the end of each trial about the percentage (proportion) of the area they explored, their average estimates were in line with the area they explored when applying the 2.5° radius UFOV to their eye movement scan paths.

The second consideration is that observers might not necessarily have explicit knowledge concerning the exact spatial extent of their target-specific empirical UFOV but rather some perceived UFOV estimate that they can utilize to compute the area explored. The perceived UFOV might not always agree with the empirical UFOV as evidenced by

peripheral inflation or filling in, a phenomenon where our metacognitive estimates of target detectability in the visual surround, away from the point of fixation, can be inflated and do not track the fidelity of information processing in the visual periphery (Odegaard et al., 2018; Solovey et al., 2015). Peripheral inflation gives rise to the rich phenomenological experience of seeing everything at once in our visual field at any point in time (Knotts et al., 2019) despite the neurophysiological and attention constraints that cap our peripheral detection capabilities (Stewart et al., 2020). Considering these two assumptions, we hypothesize that observers use the target-specific perceived UFOV, rather than target-specific empirical UFOV, to track the proportion of area already explored during their search. Moreover, an observer compares this introspective estimate to a stopping criterion to terminate their search when they fail to find the target.

Until now, our hypothesis could account for variations in search exploration across different targets. However, the hypothesis would not explain the differences in the area explored between 2D and 3D searches. Lago et al. found that regardless of the target being searched for, humans explored roughly double the area in the 2D search condition than in the 3D search condition. However, when asked to estimate the proportion of the area explored, observers massively overestimated the area explored for 3D searches. On the other hand, those reported estimates for the 3D searches were similar to the more veridical estimates of the area explored in the 2D search conditions (Lago, Jonnalagadda, et al., 2021). Those results motivate us to test a simple hypothesis. When searching through 3D image stacks, observers keep track of the proportion of area explored in the 2D image plane without regard for the area yet to be covered in the rest of the volumetric image. The second hypothesis posits that observers search approximately the same proportion of area in the 2D

search as the 2D plane of the 3D search (i.e., 3D fixations projected onto the 2D image plane) with the target-specific perceived UFOV.

To test our hypotheses, we investigated 2D and 3D searches for two targets with different UFOVs. One target is larger and more detectable in the visual periphery, while a second small target is difficult to detect away from the fovea. The purpose of Experiment 1 was to estimate for each observer a target-specific empirical UFOV by measuring the detectability (yes/no task, 50% prevalence) of the two targets at different eccentricities and polar angles (location-known-exactly to the observer). We also estimated the target-specific perceived UFOVs using two different procedures for which observers assessed target detection accuracy at different eccentricities and polar angles. The first method to measure the target-specific perceived UFOV required observers to estimate the accuracy of detecting the target in the visual periphery while fixating on the target foveally. The second method presented the target at different retinal eccentricities so that observers could experience it in the visual periphery while making perceptual judgments about their perceived accuracy for detecting it at those locations.

Experiment 2 measured visual search for the two targets in 2D displays and 3D image stacks to test our hypotheses related to common metrics used to search in 2D and 3D images of different targets. Based on this experimental setup, our hypotheses make the following predictions. First, the proportion of the area explored with the target-specific perceived UFOV (but not target-specific empirical UFOV) will be approximately equal across the small and large targets, suggesting a unitary metric being used as a stopping criterion. Our second hypothesis predicts that when observers perform the 3D search, the proportion of the 2D image plane area covered with the target-specific perceived UFOV will be similar to the

proportion of area covered with the same type of UFOV but in the 2D search condition, also suggesting a unitary stopping criterion.

2.3. Experiment 1

2.3.1. Methods

Participants

Six undergraduate students (50% female, age range: 19-22) at the University of California, Santa Barbara, participated in experiments 1 and 2 for course credit. The sample size was based on the estimated search error rate effect size from a previous experiment utilizing similar stimuli. A repeated measure t-test from Klein and Eckstein (D. Klein & Eckstein, 2023) compared the search error rate in 2D and 3D search for a small target (N=6) and found an effect size of 5.8, therefore requiring N=3 to obtain 90% power in a two-tailed t-test. Previous studies with similar sample sizes have also established a large difference in search error rate for the small target between 2D and 3D search (D. S. Klein et al., 2023; Lago, Abbey, et al., 2020; Lago et al., 2018). Our choice of six participants was also based on practical constraints. Participation in both experiments took approximately 1.5-2 months. Participants came in for 2-hour sessions three days a week.

All participants maintained normal or corrected-to-normal vision while participating, verified by the Snellen Chart for visual acuity 20/20. The University of California, Santa Barbara Institutional Review Board (IRB) approved the experimental procedures under protocol 12-22-0667. All participants signed a consent form before participating in the experiments.

Apparatus

In both experiments, participants interacted with stimuli on a medical-grade Barco MDRC-119 LCD monitor (16.5-in. x 13.5-in; screen resolution of 1,280 x 1,024 pixels; screen width 37.5cm; refresh rate of 60 Hz). The monitor was linearly calibrated for luminance intensity such that 0.1 cd/m^2 and 111 cd/m^2 corresponded to gray level values of 0 and 255, respectively. Participants sat in a darkened room (2 lux) at a viewing distance of 75 cm from the monitor, translating to 45 pixels per degree of visual angle (dva).

While performing the various tasks in each experiment, an EyeLink-1000-plus desktop mount real-time eye tracker (SR Research Inc.) monitored the participant's right eye movements at 2,000 Hz. The default parameters—eye velocity and acceleration thresholds of 30 °/sec and 9,500 °/sec^2, respectively—defined the onset of a saccade. At the beginning of each block of trials described in the methods sections below, participants completed a 9-point calibration and validation procedure to ensure valid eye-tracking data. All tasks were controlled (i.e., presentation of visual stimuli and recording of keystrokes and mouse events) in Psychopy (Peirce et al., 2019), a Python programming package for psychophysical experiments.

Stimuli

Background

To generate a 3D volumetric image, we first populated a 3D array of size 1,024 x 820 x 100 pixels with IID noise—gray levels sampled from a Normal distribution, $\mathcal{N}\sim(128, 25)$. We filtered the white-noise array to introduce pixel-to-pixel correlations amongst the gray level values in the x, y, and z dimensions. The filtering process simulates a

mammogram's idealized noise power spectrum (NPS) ($\frac{1}{f^{2.8}}$; where f denotes radial frequency index). The NPS follows a power law assuming a stationary stochastic image generation process (Abbey & Barrett, 2001; Burgess et al., 2001). A single 3D array was represented as 100 slices (i.e., 100 2D images). Each slice was of size 22.8 x 18.2 dva. The 2D noise texture background extensively used in Experiment 1 was defined as the 50th slice taken from 100 slices constituting a 3D volumetric image.

Targets

The small and large targets had two distinct geometric shapes but were matched in peak contrast, 0.43—defined as the additive luminance of the target (24.02 cd/m^2) divided by the mean luminance of the background (55.77 cd/m^2). In object space, the small target was modeled as a sphere with a radius of 2 a.u. In image space (on a 3D Cartesian coordinate system), we discretized the sphere into five circular disks, as shown in Figure 2.1.a, bottom left. Each disk represents a cross-sectional view of the sphere in the xy-plane at a different coordinate in the z dimension. Thus, in the z dimension, the sphere's diameter was five slices with a radius of 2 slices extending up/down from the central slice, denoted as slice *c*. The central slice of the sphere (Figure 2.1.a, top left) was a circular disk with a radius of 2 pixels (0.044 dva) and a diameter of 5 pixels. The two circular disks above and two below the central slice were parameterized by smaller radii such that radius $_c$ > radius $_{c\pm1}$ > radius $_{c\pm2}$. Each constituent disk of the sphere maintained a uniform contrast of 0.43.

The large target's shape (Figure 2.1.a right) was modeled by a 3D Gaussian function in object space ($\sigma = 20$ a.u.). In image space, it was parameterized by $\sigma_{xy} = 20$ pixels or 0.44 dva and $\sigma_z = 20$ slices. The large target's centroid maintained a contrast of 0.43. However,

its contrast monotonically decayed away from its center in the xy-plane. Figure 2.1.a, top right, exemplifies how the contrast decays from the center. Moreover, the peak contrast on the non-central slices (i.e., $z \neq c$) was lower than on the central slice. Therefore, the target's contrast also decreased away from the center in the z dimension. Figure 2.1.a, bottom right, shows how the peak contrast changes as a function of the distance from the central slice.

In Experiment 1, we only utilized 2D profiles of each target, which were simply the central slice of the 3D profiles (Figure 2.1.a, top row).

Final image stimulus

In order to generate a final image stimulus for each of the three tasks described below (task 1-peripheral detection, task 2-foveal perceive, and task 3-peripheral perceive), we applied the following procedure. First, we generated a 3D array and selected the 50^{th} slice out of 100 for 2D image conditions. Therefore, participants only interacted with 2D displays in Experiment 1. Target-absent stimuli in task 1 were simply this slice. For stimuli that contained a target (task 1 target-present stimuli and tasks 2 and 3), we followed the same procedure to generate a target-absent stimulus. However, we then linearly added the central slice of a single target profile (Figure 2.1.a, top) to a particular (x, y) coordinate in the image. Therefore, the center-of-mass of the target was at that position. The array elements were then converted to unsigned 8-bit integers and stored as a PNG file.

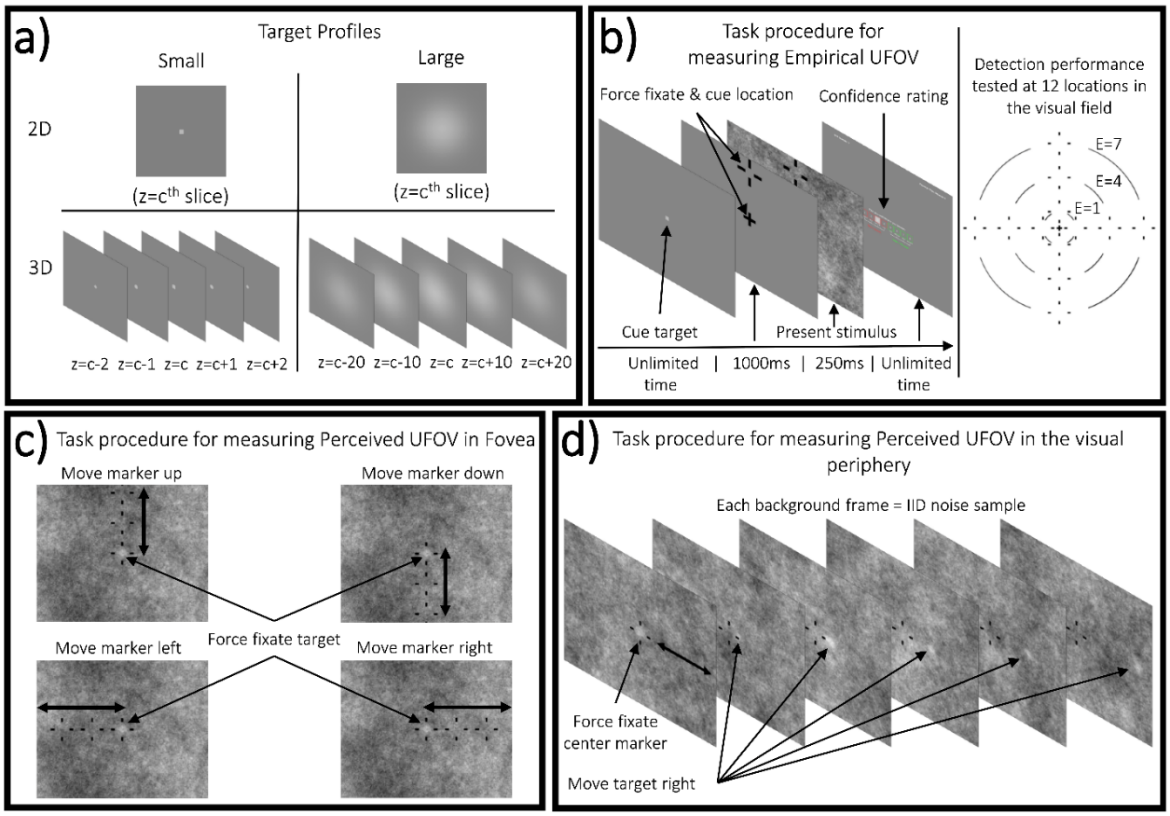

Figure 2.1. Target profiles and task procedures for measuring the empirical and two perceived UFOVs. a) Depictions of the small and large target profiles in 2D (top) and in 3D (bottom). b) A forced fixation yes/no detection task to determine target detectability as a function of retinal eccentricity was used to construct the empirical UFOV. (Left) The task procedure and the trial timeline for the detection task. (Right) The twelve possible locations in the visual field where a target could appear fall along the four cardinal directions (left, right, up, and down) at three eccentricities (1, 4, and 7 dva). c) The task procedure for measuring the perceived UFOV in the fovea. Observers estimated the peripheral detectability of the target while foveating it. Four example trials exemplify how the marker can move along the four cardinal axes. d) The task procedure for measuring the perceived UFOV in the visual periphery. Rather than having the marker move into the visual periphery like in c), the target would move into the visual periphery so that observers could experience the target's visibility in the peripheral field of view. All 2D slices shown are from a single trial. However, each slice is sampled from an independently generated 3D image.

Design and procedure

Task 1 – Measuring the target-specific empirical UFOV ($UFOV_E$)

The measurement of the target-specific empirical UFOV is based on the standard yes/no SDT experiment paradigm (Lago, Sechopoulos, et al., 2020). A noisy 2D texture is presented to the observer for a brief period. The texture contains a single target at a known location on a fraction of the trials. The observer must indicate whether the target was present

or absent at the cued location after the period has elapsed. Our task builds upon this base experimental procedure by recursively applying it with the target positioned at different retinal eccentricities and polar angles across trials to produce a performance map across the visual field (Carrasco et al., 2001; Lago, Sechopoulos, et al., 2020; Najemnik & Geisler, 2005). Figure 2.1.b, left, depicts an example trial for detecting the small target positioned in the top portion of the visual field at a distance of 7 dva from the center position of the monitor, the location where the participant was instructed maintain fixation while the image stimulus appeared on the screen.

At the beginning of each trial, a high-contrast copy of one of the two targets was presented to the participants. After clicking the space bar, participants maintained fixation on a black cross placed at the center of the screen on top of a gray background. A fiducial marker was also present at one of three retinal eccentricities (1, 4, or 7 dva) along one of the four cardinal axes. Figure 2.1.b, right, depicts the 12 possible locations where the marker could appear relative to the center fixation cross. The salient marker cued participants to the location in the visual field to attend to covertly. After maintaining fixation on the cross for 1 second, the stimulus appeared on the screen for 250 ms. The stimulus either contained the target at the center of the cued location or did not (50% prevalence). If participants broke fixation (1 dva distance tolerance), the trial would abort. Afterward, participants encountered an 8-point rating scale where they had to rate their confidence in their decision. A rating of 1-4 was reserved for target-absent decisions, with 1 representing the highest confidence that the target was absent and 4 representing the lowest confidence that the target was absent. Conversely, for target-present decisions, a rating of 8 corresponded to the

highest confidence that the target was present, and a rating of 5 mapped to the lowest confidence in their target-present decision.

This procedure was repeated 1,200 times in 50-trial block increments. Each combination of conditions (eccentricity, three levels; direction, four levels; target type, two levels) included 25 target-present trials and 25 target-absent trials. We intermixed the conditions, resulting in a random presentation order across trials.

Task 2 – Measuring the target-specific perceived UFOV with a foveal reference target ($UFOV_{P\,F}$)

In task 2, rather than having participants detect the targets in their parafovea or visual periphery, they had to estimate how well they could see the large and small targets in their visual field. Before the experiment began, participants were given the following instructions: "You will move a marker to a location on the screen where you think you would be able to achieve a performance level of, say, 70% accuracy if the target you are currently staring at was placed at that location in the noise background and you could not move your gaze from the center of the screen. The stimulus would appear for only a quarter of a second, and half the time, the target would be there at that location, and the other half would not. In other words, if there were 100 trials, you would make a correct decision on 70 trials at that location you move the marker to." Participants were allowed to ask clarifying questions after instructions. One way of thinking about this task is that observers extrapolate foveal information to the visual periphery during fixation (Stewart et al., 2020).

At the beginning of each trial, participants were presented with one of five percentage correct (PC) values: 50%, 60%, 70%, 80%, or 90%. After acknowledging the PC value, participants fixated on a black cross on a gray background positioned at the center of the

screen for one second. Afterward, a single image stimulus would appear on the screen with only one of two reference targets presented foveally where the fixation had been presented. The same marker used to cue the target locations in task 1 surrounded the target, superimposed on top of the image stimulus. Participants would then manipulate the mouse scroll wheel to move the fiducial marker to a position they estimated the foveally presented reference target would be detected with the accuracy (PC) assigned for that trial. The fiducial marker could be moved along only one of the four cardinal directions in increments of 0.111 dva. The center of the marker always started at 0 dva and could extend as far out as 1 dva from the edge of the image stimulus in a particular direction (10.4 dva from fixation horizontally and 8.1 dva vertically).

Figure 2.1.c exemplifies four separate types of trials for estimating the large target's detectability in the visual periphery. Each example image shows a subset of valid positions where the marker could land along a particular cardinal axis. After positioning the marker at a location, participants pressed the space bar to confirm. The dependent variable of interest was the distance between the marker's final position and the center of the screen where the participant was fixating. At the end of each trial, participants were presented with the 5 possible PC values that could be assigned to a trial. They had to select the correct PC value presented at the beginning of the trial. This allowed us to identify trials for which the observers misremembered the assigned PC (3% of all trials dropped because the observer reported PC at the end of the trial did not match the assigned PC). The trial quit if participants broke fixation at any point.

Participants completed 800 trials in total, 50 trials per block. There were 40 combinations of conditions (five PC values, four directions, and two targets), amounting to

20 estimates per combination. We randomized the trial presentation order across the 40 combinations for each participant.

Task 3 – Measuring the target-specific perceived UFOV with a peripheral reference target ($UFOV_{P\,P}$)

We utilized a second techniques to measure the perceived UFOV that involved the observer's access to the sensory signal of the target in the visual periphery while estimating peripheral detectability. The critical methodological difference between the foveal estimation task described above (task 2) and the peripheral estimation task was that observers moved a fiducial marker into the visual periphery in the former. In contrast, in the latter, they moved the target. To this effect, in the peripheral UFOV estimation task, participants experienced the target in the visual periphery while they estimated how well they could see it. In the previous task, they were continually fixating on the target while making estimates.

The procedure's beginning was the same as task 2 (i.e., PC value first presented to an observer, fixation cross procedure, etc.). Once the stimulus appeared on the screen, the target was initially positioned where the participants were fixating. The fiducial marker surrounded the target but did not move. The leftmost image in Figure 2.1.d, without the double-headed arrow, depicts what participants first saw when the image stimulus appeared on the screen. The double-headed arrow indicates that a participant must estimate the large target's detectability in the right portion of the visual field on that particular trial. Specifically, while maintaining fixation at the center of the screen, the participant would manipulate the mouse scroll wheel to move the target in increments of 0.44 dva along the right cardinal axis to the edge of the image stimulus. The max distance from the center of

the screen to where the center of the target could be placed was 8.4 dva to prevent portions of the large target from being cropped out of the image. We chose to keep this distance the same across target types.

Each time the participant moves the target, a different (IID) noise texture background appears on the screen, with the target embedded in a more distal (or proximal) location relative to the fixation position. The five image stimuli on the righthand side of Figure 2.1.d demonstrate how the target appeared at five more distal locations along the right cardinal axis. Including different 2D noise images (i.e., the 50^{th} slice from independently generated 3D images) for each scroll event mitigated the effect of a motion-percept confound whereby the target is perceived as moving across a fixed background.

Once participants placed the target at a desired location in their visual periphery, they pressed the spacebar to end the trial and then entered which proportion correct value (e.g., 50%, 60%, etc.) they were prompted with at the beginning of the trial (2% of all trials were discarded because the reported PC did not agree with the prompted PC at the beginning of the trial). Broken fixations led to the early termination of the trial.

This procedure was repeated ten times for each direction (four levels: up, down, left, and right), proportion correct estimate (5 levels), and target type (2 levels) combination, totaling 400 trials. The task was broken up into eight 50-trial blocks. We chose to halve the number of estimates here relative to task 2 because rendering the image stimuli onto the graphics card for a single trial took quite some time. Participants completed thousands of trials across experiments 1 and 2, and we did not want to overburden them. We also ran a power analysis (Table A.1) on the estimates from task 2 to demonstrate that ten estimates per combination of conditions are sufficient.

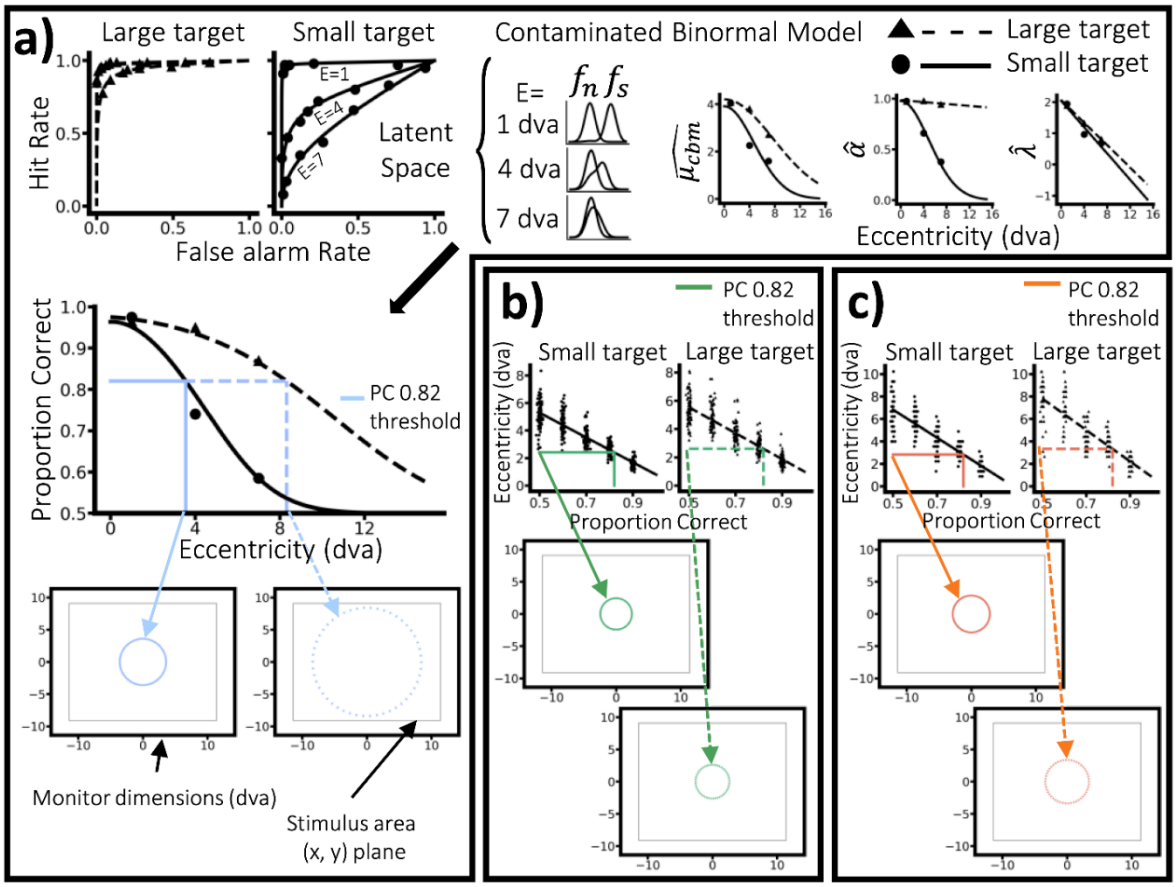

Figure 2.2. Algorithm for generating different types of UFOVs. a) (Top left) For a given subject and both targets, Contaminated Binormal ROC curves are fit to rating data from the yes/no detection task (task 1) at three different retinal eccentricities (1, 4, 7 dva). (Top middle) The latent space of target detectability for the small target at all eccentricities is conceptualized by a signal distribution, noise distribution, and a criterion, λ (not shown). Moreover, the signal distribution is defined by a recentering parameter, μ_{cbm}, and a mixing parameter, α. (Top right) For each eccentricity, three parameters are estimated. For each parameter, secondary fits (half Gaussian or linear) are applied to the three point estimates. (Middle left) Together, these parameters and secondary fits are used to estimate the proportion correct for untested eccentricities. For the proportion correct threshold of 0.82 (blue line), an eccentricity is estimated for the large and small targets. (Bottom) Those two eccentricities (arrows pointing to circles) are the radii of target-specific empirical-UFOVs, $UFOV_E$. The size of the UFOVs is shown in relation to the stimulus and monitor dimensions. b) For the large and small targets, a linear prediction of eccentricity at the proportion correct threshold of 0.82 (green line) is used for the radii (arrows pointing from the y-axis to circles) of the target-specific perceived UFOVs from the foveal estimation task (task 2), $UFOV_{PF}$. Note, random noise is added to the x position of each scatter point for display purposes only. c) For the peripheral estimation task (task 3), the same linear fit procedure as shown in b) is done for computing the radii of the second type of Perceived UFOVs (orange line), $UFOV_{PP}$.

Fitting individual target-specific empirical and perceived UFOV

The main objective of tasks 1, 2, and 3 was to derive target-specific empirical and perceived UFOVs for each participant. We first selected a proportion correct threshold of

0.82. This means that a target appearing anywhere within the circumscribed region of the UFOV, which is centered on a fixation point, has a probability of being detected at 0.82 or greater. A proportion correct of 0.82 is common in the psychophysics threshold detection literature (Britten et al., 1992; Cameron et al., 2002; Najemnik & Geisler, 2005). By fixing this free parameter, we could normalize the spatial extent (or area) of the empirical UFOV with the areas of the two perceived UFOVs. However, 0.82 is not a level of our independent variable in task 2 or 3, nor is there an eccentricity that produced a PC of 0.82 in task 1. Therefore, we needed to estimate from our data the eccentricity that produces a PC of 0.82 in task 1 and determine what eccentricity is predicted for a PC of 0.82 in the latter two tasks. Below, we describe these methods.

The empirical UFOV ($UFOV_E$) was obtained for a given participant and target type via multiple fits. First, we fit a contaminated binormal model (CBM) (Dorfman & Berbaum, 2000) to the rating data obtained in task 1 for each target at every eccentricity (100 target-absent ratings and 100 target-present ratings after collapsing across the four cardinal directions). We chose the CBM as it is robust in estimating proper ROC curves—curves that do not hook under the chance line. Figure 2.2.a, top left, depicts the fitted ROC curves for the two targets at each eccentricity for a single participant using maximum likelihood estimation.

The CBM estimates two parameters of the well-known signal distribution, which resides on a latent axis in the Signal Detection Theory literature (Green, 1966; Macmillan & Creelman, 2005). Rather than treating the signal distribution as a shifted standard normal (assuming equal variance), the CBM assumes the signal distribution is comprised of a mixture of Gaussians—the first being the standard normal and the second being a unit

variance Gaussian centered on μ_{cbm}, where $\mu_{cbm} \geq 0$. The second parameter, α, denotes the mixing fraction of the two distributions and is bounded between 0 and 1. Figure 2.2.a, top middle exemplifies the latent distributions for the small target at the three eccentricities based on the CBM fits. Lastly, the CBM estimates K-1 cut points, where K is the number of rating options used in the experiment. We chose to focus on the middle cut point (K=4), which coincides with the criterion, λ, that separates target-present decisions (ratings greater than 5) from target-absent decisions (ratings less than 5) in our experimental paradigm. The CBM assumes a standard normal for the noise distribution. For a given combination of parameter estimates, $\widehat{\mu_{cbm}}, \hat{\alpha}$, and $\hat{\lambda}$ we computed an estimated PC with the following set of equations:

$$\widehat{TPR} = (1 - \hat{\alpha}) * \phi(-\hat{\lambda}) + \hat{\alpha} * \phi(\widehat{\mu_{cbm}} - \hat{\lambda}) \qquad (Eq. 2.1)$$

$$\widehat{FPR} = \phi(-\hat{\lambda}) \qquad (Eq. 2.2)$$

$$\widehat{PC} = \frac{\widehat{TPR} + (1 - \widehat{FPR})}{2} \qquad (Eq. 2.3)$$

Where ϕ denotes the cumulative distribution function of the standard normal distribution.

We measured observer performance at three eccentricities: 1, 4, and 7 dva. Figure 2.2.a, top right, depicts the relationship between our parameter estimates, $\{\widehat{\mu_{cbm}}, \hat{\alpha}, \hat{\lambda}\}$ and visual eccentricity for one observer. However, we required PC estimates for untested eccentricities. To achieve this objective, we fit a half Gaussian function to the parameter estimates of $\widehat{\mu_{cbm}}$ for both targets. The half Gaussian asymptotes at 0 as eccentricity increases, which is the lower constraint on the bounds for this parameter. For the small target, $\hat{\alpha}$ was also fit to a half Gaussian to avoid negative values. For the large target $\hat{\alpha}$ estimate, we fit a line. We also fit a line for the $\hat{\lambda}$ estimates for both targets. We utilized

these secondary fits to generate predictions of PC for untested eccentricities using *Eq. 2.1-2.3*, as shown in Figure 2.2.a, middle left. Lastly, we determined the eccentricity that produced an estimated PC of 0.82, which, in turn, defines the radius of the circular $UFOV_E$ for each target, as shown in blue at the bottom of Figure 2.2.a.

The radii for the two types of circular target-specific perceived UFOVs were computed via simple linear fits to the raw data obtained from the foveal estimation task (task 2-$UFOV_{P\,F}$) and the peripheral estimation task (task 3-$UFOV_{P\,P}$). Once again, we collapsed our data across directions before regressing eccentricity on PC, as depicted in the top of Figures 2.2.b and 2.2.c. We then predicted the eccentricity for a PC of 0.82. These eccentricities served as the radii of the UFOVs (Figures 2.2.b and 2.2.c bottom). Individual fits for each subject can be found in Figure A.1.

Data analysis

Statistical analyses

Our first analysis considered the slope and intercept estimates for a simple linear regression model that predicted proportion correct from eccentricity (task 1). For tasks 2 and 3 the linear models predicted eccentricity from the probed proportion correct. For each task, our goal was to assess differences in linear fits between targets at a participant-average level. This allowed us to identify the differences in peripheral detectability of each target and differences in metacognitive estimates regarding peripheral detectability. Our second analysis focuses on individual differences across the three tasks to derive target-specific empirical and perceived UFOVs for each participant. These individualized target-specific UFOVs are pertinent for our analyses in Experiment 2.

Statistical analyses used non-parametric bootstrap resampling procedure of trials and participants, with replacement, to construct empirical sampling distributions (n=20,000) of the slope and y-intercept estimates from the data collected in tasks 1-3. For example, let us consider one bootstrap iteration for the data collected in task 1. After sampling with replacement trials and participant IDs, we computed the proportion correct at each of the three eccentricities for the two targets (6 PCs in total). We fit a line to the 3 PC estimates for the small target. The same procedure was done for the large target.

Next, we computed the difference between the slope estimates of the two targets for each bootstrap iteration. Out of the distribution of 20,000 difference scores, we computed the fraction of scores that were less than 0. We multiplied this fraction by 2 to obtain an unadjusted two-tailed p-value. This procedure was repeated for the y-intercept as well. Finally, we applied an FDR correction (Benjamini & Hochberg, 1995) to the p-values based on the six comparisons tested in Experiments 1—2 for each task.

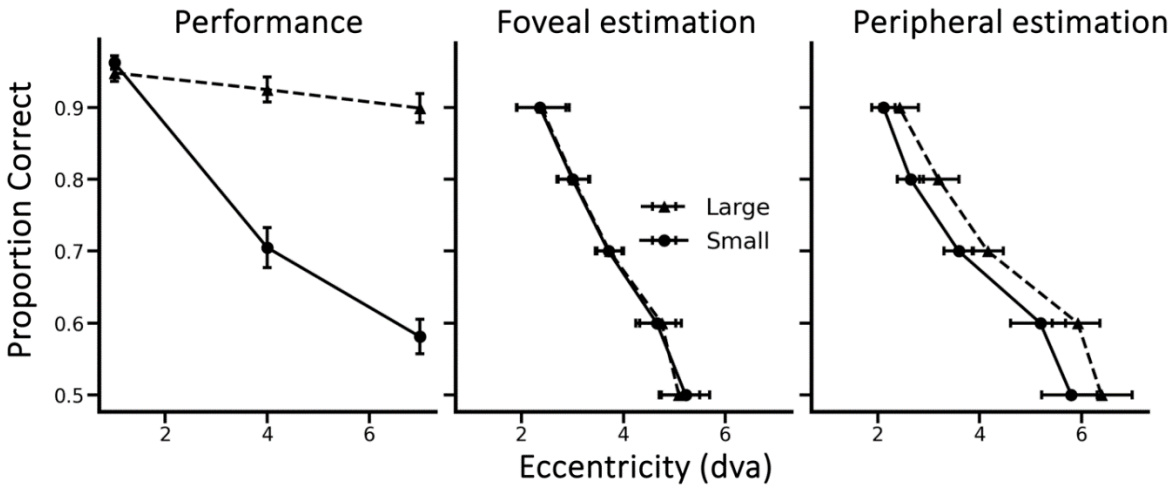

Figure 2.3. Proportion correct plotted versus eccentricity for the detection task (task 1) and two estimation tasks (tasks 2 and 3). (Left) Mean detection performance is plotted at eccentricities 1, 4, and 7 dva (collapsed across polar angle). (Middle) For the following proportion correct thresholds: 0.5, 0.6, 0.7, 0.8, and 0.9, mean eccentricity across observers is plotted for the foveal estimation task. (Right) It is the same as the middle plot but for the peripheral estimation task. Triangle points represent the mean performance across observers for the large targets, and circular points represent the mean performance for the small target. Error bars represent 68% confidence intervals (~1 standard error) from bootstrap sampling distributions.

2.3.2. Results

Actual versus estimated target detectability across the visual field

In task 1, we found a significant difference in the slope estimates between the large and small targets ($\Delta\beta_1 = 0.0554$, $p < 5e^{-5}$). Figure 2.3, left, shows a steep drop-off in performance as eccentricity increases for the small target relative to the large target. Although the two targets' peak contrasts are matched, low visual acuity and spatial resolution in the periphery hindered the detection of the small target, comprised of high spatial frequency information. We also found a significant difference in the y-intercept ($\Delta\beta_0 = -0.0469$, $p = 0.0087$).

In task 2, we found no significant difference in either the slope ($\Delta\beta_1 = 0.0026$, $p = 0.6065$) or y-intercept ($\Delta\beta_0 = -0.0866$, $p = 0.6591$). When estimating target detectability in the fovea, on average, participants considered both targets equally detectable in their peripheral field of view (Figure 2.3, middle). However, in task 3, despite not finding a significant difference in the slope ($\Delta\beta_1 = -0.0075$, $p = 0.5675$), we found a large difference in the y-intercept, although it was also not significant ($\Delta\beta_0 = 1.0785$, $p = 0.1242$). Figure 2.2, right, depicts a rightward shift in the large target estimates, suggesting that the participants may think they see the large target better in their visual periphery than the small target. Figure A.2 breaks down these analyses by direction.

2.3.3. Discussion

In this experiment, we have demonstrated how the observer's perceived concerning target detectability in the visual periphery differs from the empirical peripheral detectability. Specifically, participants misjudge that both targets are equally detectable at large eccentricity. However, detection performance for the small target is significantly worse at 4

and 7 dva than for the large target. Second, we have utilized the measured and estimated target detectability across the visual field to construct three types of UFOVs at a fixed proportion correct threshold of 0.82. These UFOVs will be a tool to compute the proportion of area explored in the 2D and 3D searches described in the following experiment.

2.4. Experiment 2

2.4.1. Methods

Participants and apparatus parameters were kept constant between experiments 1 and 2.

Stimuli

The stimuli used for the 2D search task were generated in the same manner as in Experiment 1, task 1 (see “Final image stimulus” subsection in the methods section of Experiment 1) but with one important caveat. The target's center (x, y) position was randomly generated for each target-present stimulus. Specifically, we uniformly sampled an x and a y coordinate within the confines of a rectangular area that was smaller than the image stimulus dimensions and then placed a single target (Figure 2.1.a, top-left or top-right) at that location to avoid cropping the target’s edges by the image stimulus boundaries. In total, there were 200 independently generated stimuli: 50 stimuli contained the large target, 50 stimuli contained the small target, and 100 stimuli contained neither target.

The stimulus set for the 3D search task consisted of 200 independently generated 3D volumetric images. Recall from the Methods section of Experiment 1 that a single volumetric image comprised 100 2D slices. Therefore, participants could view all 100 slices of the 3D image in a single trial of the 3D search task. One hundred 3D volumetric images

contained no target, fifty 3D volumetric images contained a small target, and the other fifty 3D backgrounds contained a large target.

In order to embed a single 3D target into the 3D background, we applied the following procedure. First, we randomly sampled an (x, y, z) coordinate, denoted as (x^*, y^*, z^*). The central slice of a single target (e.g., Figure 2.1.a, bottom-left, slice z=c) was then linearly added to the z^* slice of the 3D background at the (x^*, y^*) position. The slices of the target above and below the central slice (e.g., z=c + 1 and z=c – 1 in Figure 2.1.a, bottom-left) were added to the z^* + 1 and z^* – 1 slices of the 3D image at the same (x^*, y^*) position. This procedure was repeated until all slices of a given target were added to the 3D background.

The location (x^*, y^*, z^*) was constrained to be within a smaller cube inside the volumetric image to prevent cropping of the target profile by the boundaries of the image stimulus. Like the 2D target-present stimuli, (x^*, y^*) was sampled from a rectangular area smaller than the area of a single 2D slice. In the third spatial dimension, we ensured that z^* was never in slices 1-10 or 90-100. Therefore, all five slices appeared in the 3D volume for the small target. For the large target, there were instances where not all slices of the 3D profile were inserted into the 3D background. For example, in the edge case that $z^* = 11$, the target profile slices greater than c+10 would not be inserted into the 3D volumetric image.

Design and procedure

2D search

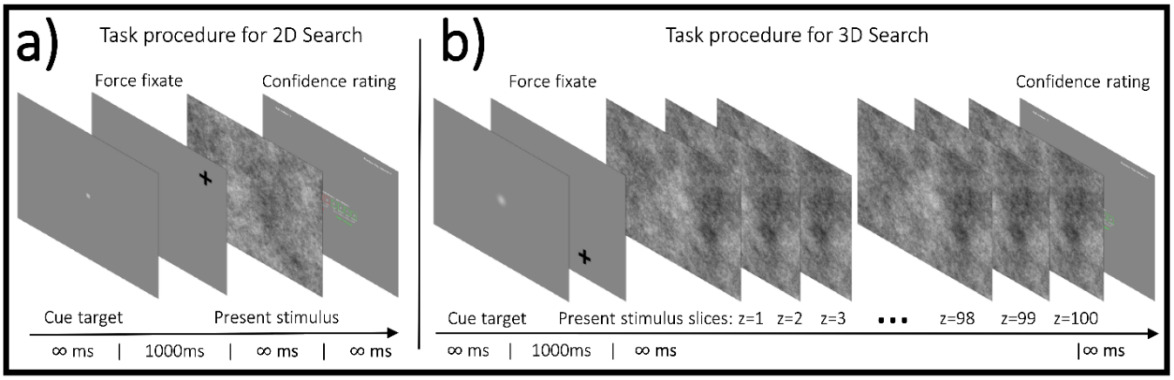

Figure 2.4. Depiction of the trial flow for the 2D and 3D searches. a) The 2D search task procedure is depicted. At the onset of the trial, the target type is cued (small or large). Next, a participant must fixate on a randomly located cross for one second. Afterward, the visual stimulus would appear. The participant would search for the cued target in the image until they find it or decide to quit the search. Lastly, they provided a confidence rating in their decision. b) The 3D search task procedure is depicted. It follows the same general structure as the 2D search task. However, participants can scroll through 100 slices (z=1, z=2, etc.) constituting the 3D volumetric image.

Participants completed 200 trials in the 2D search task. We intermixed the target-present and target-absent stimuli for both types of targets and randomized the presentation order. In other words, an observer could see a small-present, small-absent, large-present, or large-absent stimulus on any given trial. Participants completed four 50-trial blocks in total, with breaks in between blocks.

Before each trial began, participants were informed about which target to look for with the presentation of a reference target (Figure 2.4.a, left). Next, they were instructed to stare at a fixation cross (Figure 2.4.a, middle left), randomly located within the image stimulus boundaries, for one second to ensure proper eye tracker calibration (i.e., a custom drift check). Participants were able to start a complete recalibration at this time if needed. Afterward, the image stimulus appeared on the screen (Figure 2.4.a, middle right), and participants were allowed to make free eye movements. They had unlimited time to perform the search. A high-contrast copy of the 2D target profile (Figure 2.1.a, top row) was placed above the image stimulus to remind participants which target they had to look for.

To end the trial, participants had to make one of two choices: press the spacebar to end the trial to signify a target-absent decision or click at a location where they thought the target might be present in the image and then hit the spacebar key to confirm (both actions together denoting a target-present decision). Afterward, they were instructed to rate their confidence in their decision on an 8-point scale (Figure 2.4.a, right), the same rating scale used in Task 1 of Experiment 1.

No feedback was given at the end of the trial. However, participants completed practice trials with feedback before starting the 2D search task to help familiarize themselves with the task. The feedback on target-present trials consisted of displaying the image stimulus again after participants input a confidence rating. A circular ring was superimposed on top of the image stimulus around the target location to demarcate its position. In the target-absent trials, a gray background with the text "ABSENT" was presented to the participants. In both instances, there was unlimited time to review the feedback.

3D search

Participants searched through 200 3D volumetric images (50 % prevalence). They saw one volumetric image per trial. In one hundred trials, participants were asked to search for the small target; in the other half, they were asked to look for the large target. Given the relatively long nature of the search, the trials were broken up into sets of 10 trials per block and randomized across the different target types and ground truth statuses, like in the 2D search task.

The basic procedure/trial flow of the 3D search task (Figure 2.4.b) is akin to that of the 2D search. Participants were instructed to look for a particular target at the beginning of the trial, and a fixation cross appeared at a random location on top of a gray background within

the confines of the (x, y) plane of the image stimulus. After maintaining fixation at the cross for one second, the stimulus presentation portion of the trial began. Rather than seeing a single image on the screen, as in the 2D search task, participants could scroll through 100 images, each representing a different planar view of the 3D volume (Figure 2.4.b, middle). Only one image would appear on the screen at a time. However, participants could either manipulate a mouse scroll wheel to view different slices of the 3D volume or click, hold, and drag a custom scrollbar widget on the right-hand side of the stimulus to maneuver through the third dimension. The custom scrollbar also allowed participants to jump across multiple slices at a time.

To make a localization decision in 3D, participants were instructed to click on the image stimulus on the slice where they first detected the target, as the targets spanned multiple slices. Clicking on the screen produced a red circle to demarcate the localized region. Participants were instructed to make only one click per trial. To end the trial, participants clicked the spacebar button and encountered the same 8-point rating scale as in the 2D search task. The trial would conclude after they made a rating decision.

Participants completed practice trials at the beginning of the 3D task. These trials contained feedback regarding the presence/absence of the target. If the trial was target-present, the slice of the 3D volume containing the central slice of the 3D target was shown to the observer. A white circle demarcated the target location. We chose to display the slice of the 3D volume that contained the target's central slice because the target's central slice provided the strongest reinforcement of the target’s appearance in the background noise. We acknowledge that learning may have changed if we showed the slice they clicked on during

the trial. If the trial was target-absent, no image would appear, and the word “ABSENT” would be displayed on the screen instead.

Data analysis

Our first analysis focuses on behavioral performance measures for the small and large targets in the 2D and 3D search conditions. We evaluate the empirical area under the ROC curve, hit rate (and hit rate localized), false alarm rate, and search error rate. We supplemented the behavioral performance analysis with measures of search time and the number of eye movements executed during the searches. Together, these results will support the claim that observers under-explore the 3D volumetric images and change their search patterns for each target.

Our second analysis focuses on the search-termination criterion and its relation to the area explored with the target-specific perceived UFOV. We focus on trials where participants reported the targets as absent (misses and correct rejections) to parse a self-terminating search strategy from a quitting threshold independent of target detection.

Search performance measures characterizing 2D vs. 3D

We used various figures of merit to describe search performance for the small and large targets in the 2D and 3D searches. The AUC is a criterion-free assessment of search performance. We constructed empirical ROC curves with the rating data and then computed the area under the empirical ROC curve using the trapezoidal rule (Macmillan & Creelman, 2005). We computed the empirical AUC based on 100 trials (50 % prevalence) for each participant, search condition, and target type. We examined the hit and false alarm rates using the same data stratification. We defined hits as a rating greater than 4 on target-present trials. Similarly, we defined false alarms as ratings greater than 4 but on target-absent trials.

Both counts of hits and false alarms were divided by the number of target-present and target-absent trials, respectively, to produce a hit rate and false alarm rate.

Given that we asked participants to localize the targets in the 2D and 3D searches, we included an analysis of the hit rate localized. A target was localized in 2D if an observer produced a rating greater than 4 and clicked within a distance of 1 dva from the target's center (x, y) position. In 3D, we augmented the definition of localizing the target because the target profile spanned multiple slices. A valid localization in 3D required that an observer meet the first condition described above and that their click occurred on a slice that was less than or equal to N slices from the central slice of the target. We set N=2 for the small target because it spanned five slices in 3D. For instance, consider an observer looking for the small target inserted at location (x^*, y^*, z^*=45). Any click within 1 dva from the (x, y) location on any slice between 43-47 would be considered a localized hit.

For the large target, N was set to 23. Given that the large target's peak contrast decayed monotonically as a function of distance from the central slice, it is unclear what N should be. Therefore, we conducted a control experiment with five new observers to determine N for the large target. In this task, participants saw 100 3D volumetric images. We inserted half of the mass's 3D profile into the 3D images such that the central slice was on slice 1, the c-1 slice was on slice 2, the c-2 slice was on slice 3, etc. Participants were informed that the target was placed at the center of each 2D image slice and that the central slice of the target was present on the first slice they saw. Participants were instructed to scroll downward until they no longer saw the mass target, a form of method of adjustment. They were instructed to press the spacebar on the slice where they could not discern the signal from the background noise. The dependent variable of interest was the slice number or distance from the central

slice in z. Across the 100 trials, we computed the median slice distance from the central slice for each observer. We averaged these estimates across the five observers. Based on this supplementary experiment (Figure A.3), we set N=23.

Gaze-dependent errors analysis

We also considered eye movement patterns to bolster our analysis further. Search errors are a common metric for assessing human performance in complex decision-making tasks when eye-tracking data is available (Krupinski, 2010; Kundel et al., 1978). A search error occurred in a trial where a participant reported a false-negative decision. Upon further analysis of their fixation positions, it was revealed that they never stared directly at the target (i.e., never foveated it). In our 2D search experiment, we quantified a failure to foveate the target as an absence of fixation locations within 2 dva from the target's center (x, y) position. For the 3D search task, we built upon this definition by adding the constraint that the fixations needed to appear outside of N slices above or below the central slice of the target. For the small target, we set N=2, and for the large target, we used N=23. Thus, we defined the search error rate as the proportion of target-present trials where participants reported the target as absent and failed to foveate it.

Across the five dependent variables, we evaluated performance differences across 2D and 3D for each target separately. We applied the same bootstrap resampling procedure and non-parametric significance testing as described in the Data Analysis section of Experiment 1. In total, we assessed 8 FDR corrected p-values (we did not FDR correct for hit rate localized because it supplemented the hit rate metric).

Search time and number of fixations

Two additional measures, search time and number of fixations, were evaluated to understand better how the observer's search patterns changed across the two modalities for

the two targets. For a given search condition (e.g., 2D small target), we computed the mean time exploring on trials where observers reported the target as absent (misses and correct rejections). Similarly, we computed the mean number of fixations across all trials where observers reported the target as absent.

These two variables are correlated with the area explored with the UFOV, the dependent variable utilized to test our hypotheses. Therefore, we assessed statistically significant differences across targets for a given modality (e.g., mean time searching for the large target in 2D versus mean time searching for the small target in 2D) and across modalities for a given target (e.g., the mean number of fixations in small target 2D search versus the mean number of fixations in the small target 3D search). Adding pairwise comparisons between targets for these two dependent variables will help facilitate the discussion surrounding our main hypotheses. We computed four pairwise differences per endpoint, and FDR corrected the p-values for each dependent measure separately.

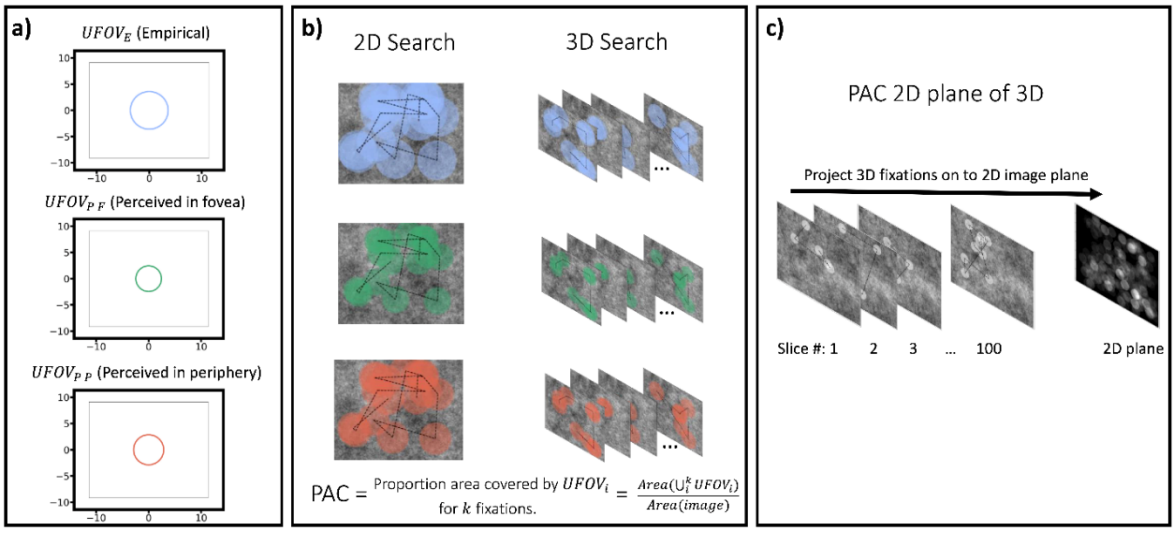

Figure 2.5. Illustration of how to compute the proportion of area explored (PAC) in 2D, 3D, and the 2D plane during the 3D search. a) The small target empirical UFOV (top), perceived UFOV in the fovea (middle), and perceived UFOV in the periphery (bottom) for a single observer are replotted from Figure 2.2. b) Graphical depiction of the search area covered in both the 2D (left column) and 3D (right column) search trials using the three types of UFOVs. Dotted lines indicate the observer's eye movement scan path in 2D and 3D, respectively. The UFOVs are painted onto the search array at each recorded fixation position during the search. The formula for computing the PAC is shown at the bottom. c) Algorithm for computing PAC of

the 2D image plane during the 3D search with an arbitrary UFOV. All fixations are projected onto one plane, and the PAC is computed as if it were a 2D search (b, left column).

Computing the proportion of area explored (PAC)

This study aimed to determine whether the proportion of area covered (PAC) by the UFOV serves as a plausible search-termination criterion in trials where observers report "target-absent." In our analyses below, which are focused on our primary hypotheses, we considered four types of UFOVs. The first is the standard UFOV with a radius of 2.5° ($UFOV_S$), which serves as a control and helps contextualize our results with what has been reported in the literature using the same UFOV size. The latter three types of UFOVS are the empirical UFOV ($UFOV_E$), the perceived UFOV in the fovea ($UFOV_{PF}$), and the perceived UFOV in the visual periphery ($UFOV_{PP}$). The spatial extent of these types of UFOVs are target-specific, whereas the standard UFOV covers the same area for both types of targets.

Figure 2.5.a illustrates an example of the three types of small target UFOVs for one observer. Figure 2.5.b illustrates how we calculated the PAC in 2D and 3D with each type of UFOV in Figure 2.5.a. In the 2D search, we painted each type of UFOV on all recorded fixation positions during the trial and determined the union set of pixels that were "painted." For instance, in the left column of Figure 2.5.b, from top to bottom, we counted the pixels colored blue, green, and red, respectively. We divided these three counts, one for each type of UFOV, by the total number of pixels in the 2D array to obtain three separate PAC values. In 3D, we also painted the UFOVs on the fixation positions on each slice visited by the observer, as shown in the right column of Figure 2.5.b. We divided this count by the total number of pixels in all 100 slices of the 3D array to compute the PAC with the UFOVs.

Our first hypothesis posits that the PAC with the perceived UFOV, but not the empirical UFOVs or standard UFOV, will be approximately equal between targets. In the 2D search, we evaluated the difference in the PAC between the two targets using each type of UFOV. We repeated this analysis for the 3D search (8 pairwise comparisons in total across the two image modalities). Next, we computed the PAC ratio for each of the four types of UFOV separately. The PAC ratio is defined as the PAC for the large target divided by the PAC for the small target. The PAC ratio facilitated pairwise comparisons between UFOV types while accounting for differences in the PAC between targets. In 2D, there were six comparisons (four UFOV types chose two). This was true in 3D as well.

We evaluated twenty pairwise comparisons for completeness, which we report in Table A.2. We used the same bootstrapping procedure discussed in the previous analyses and FDR correct for twenty comparisons. The reader can refer to Table A.2 for comparisons not directly mentioned in the Results section below.

Our second hypothesis concerns the difference in PAC between 2D and 3D searches for a given target. Here, we argue that the observers under-explore in 3D because they terminate their search after sufficiently covering the 2D image plane area with their perceived UFOV. We define *sufficient coverage* as the average PAC while performing the 2D search task for the same target. To compute the proportion of the 2D plane covered by a UFOV, we projected all the fixations in a 3D search trial onto a single 2D array (Figure 2.5.c). We then computed the proportion of the 2D array covered by the UFOVs in the same manner as Figure 2.5.b, left column.

We utilized the same bootstrap resampling procedure for each UFOV type to assess three difference scores for the large target and three for the small target. The first difference

score concerned the PAC in 2D versus the PAC in the 2D plane of the 3D search, which is our main focus. The second comparison looked at the PAC in 2D versus the PAC in 3D. The third type of comparison evaluated the difference between the two PAC ratios. The first PAC ratio was the mean PAC in 2D divided by the mean PAC in the 2D plane of the 3D search. The second PAC ratio was the mean PAC in 2D divided by the mean PAC in 3D. We evaluated twenty-four pairwise difference scores across the two targets and four UFOV types and applied an FDR correction. Results for all 24 comparisons can be found in Table A.2.

2.4.2. Results

Quantifying observer performance in 2D vs. 3D search for the two targets

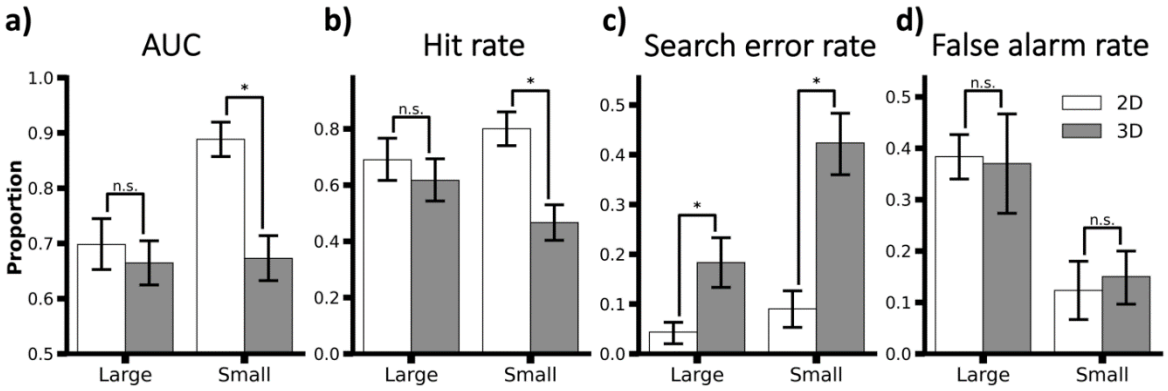

Figure 2.6. Behavioral performance for small and large targets in 2D and 3D searches. (a) AUC or area under the empirical ROC curve is depicted for the large (left) and small (right) targets. The white bars denote AUC for the 2D search condition, whereas the gray bars represent AUC in the 3D search condition. (b-c) The same stratification of data as (a) but for hit rate (b), search error rate (c), and false alarm rate (d) measures. All error bars represent 68% bootstrap confidence intervals. "*" means FDR corrected p-value < alpha level of 0.05, and "n.s." represents non-significant results.

Figure 2.6 shows the performance differences across the two image modalities stratified by the target type. The mean AUC and mean hit rate for the smaller target were significantly higher in the 2D search relative to the 3D search condition ($\bar{\Delta}AUC = 0.2156, p < 5e^{-5}$, $\bar{\Delta}HR = 0.3333, p < 5e^{-5}$). The change in hit rate localized ($\bar{\Delta}HR_{localized} = 0.4333, p < 5e^{-5}$) was also significantly different across the two searches (Figure A.4, right). As

expected, the search error rate was significantly higher in the 3D condition than in the 2D condition ($\bar{\Delta}SER = 0.36, p < 5e^{-5}$). Lastly, we did not find a significant difference in the false alarm rate between 2D and 3D searches for the small target ($\bar{\Delta}FAR = 0.0267, p = 0.3602$). These behavioral differences are similar to those found in previous studies and consistent with under-exploration of the 3D volumetric images (Lago, Jonnalagadda, et al., 2021).

The behavioral performance of the larger target tells a different story. As shown in Figure 2.6 a, b, and d, there was no significant difference in AUC ($\bar{\Delta}AUC = 0.0334, p = 0.3316$), hit rate ($\bar{\Delta}HR = 0.0733, p = 0.0771$), or false alarm rate ($\bar{\Delta}FAR = 0.0133, p = 0.7075$) when participants searched for the large target in the 2D versus 3D conditions. We observed no significant difference in the localized hit rate (Figure A.4, left, $\bar{\Delta}HR_{localized} = 0.0233, p = 0.5189$). Interestingly, the search error rate (SER) was significantly higher in 3D than in 2D for the large target ($\bar{\Delta}SER = 0.11, p < 5e^{-5}$), but to a smaller extent than for the small target.

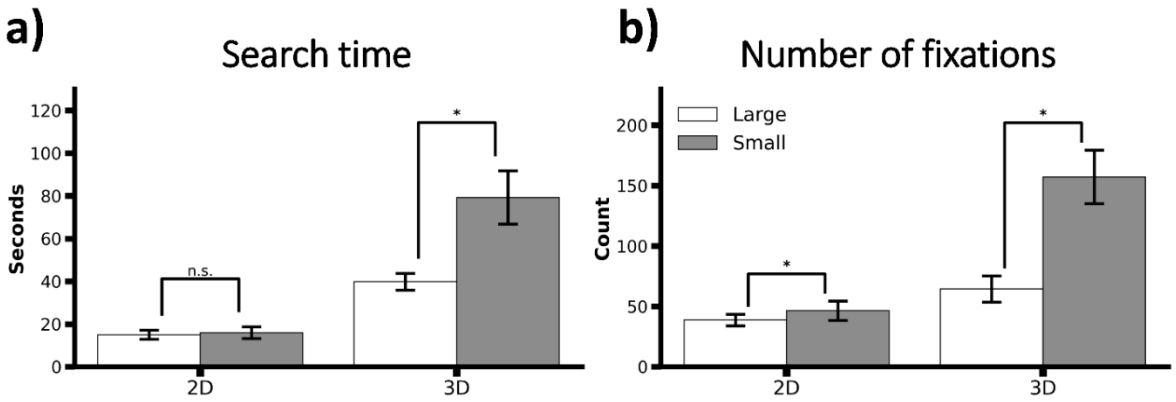

Figure 2.7. Search time and number of fixations on trials where participants reported target-absent. a) The mean search time across observers in the 2D condition (left) and the mean search time in the 3D condition (right). White bars represent the search time for the large target, and gray bars represent the search time for the small target. b) The number of fixations is plotted in the same manner as (a). All error bars represent 68% bootstrap confidence intervals. "*" means FDR corrected p-value < alpha level of 0.05, and "n.s." represents non-significant results.

Search times

In addition to the performance measures, the analysis of search time and number of eye movements shed additional light on how observers executed the 2D and 3D searches for the two targets. Figure 2.7.a shows that participants spent significantly more time in 3D searching for the small target than in 2D ($\bar{\Delta}Time = 63.2393$ seconds, $p < 5e^{-5}$). Observers also spent more time looking for the large target in 3D than in 2D ($\bar{\Delta}Time = 24.8645$ seconds, $p < 5e^{-5}$). In comparing search times across targets in 2D, there was no significant difference ($\bar{\Delta}Time = 0.9445$ seconds, $p = 0.1774$). However, in 3D, observers spent more time searching for the small target than the large target ($\bar{\Delta}Time = 39.3193$ seconds, $p < 5e^{-5}$).

Number of fixations

Figure 2.7.b conveys the mean number of fixations in the 2D and 3D searches for both targets on trials where participants reported "target-absent." Like search time, participants made significantly more fixations when searching in 3D for the small target than in 2D ($\bar{\Delta}Fix = 110.6234$ fixations, $p < 5e^{-5}$), and this pattern held for the large target as well ($\bar{\Delta}Fix = 25.8553$ fixations, $p < 5e^{-5}$). Additionally, participants made more fixations for the small target than the large target in both the 2D search condition ($\bar{\Delta}Fix = 7.7863$ fixations, $p < 5e^{-5}$) and the 3D search condition ($\bar{\Delta}Fix = 92.5544$ fixations, $p < 5e^{-5}$).

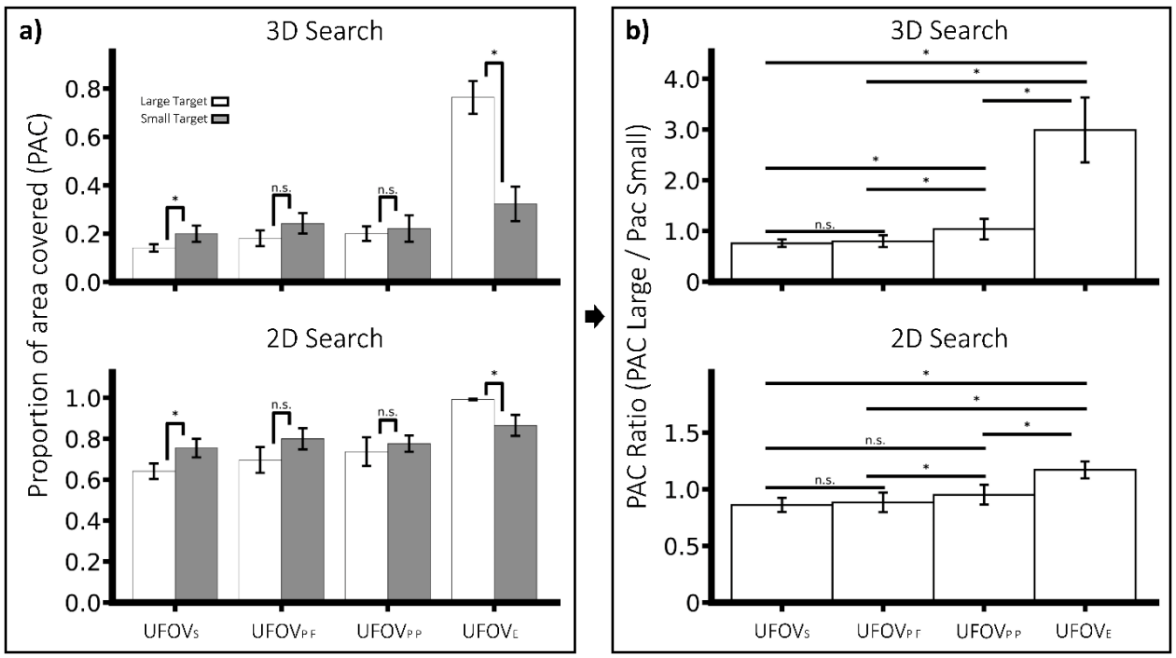

Figure 2.8. Comparing the proportion of the search area covered for the two targets with different types of UFOVs. a) The participant-average PAC in the 3D search (top) and 2D search (bottom) conditions. Gray bars correspond to the mean PAC for the small target search on false-negative and true-negative trials. White bars represent the same information but for the larger target. We include the standard UFOV in the left column for reference. From left to right, the x-axis labels are as follows: UFOV_{S}, UFOV_{PF}, UFOV_{PP}, and UFOV_{E}. b) The PAC ratio, or the large target PAC divided by the small target PAC for the 4 UFOV types in c). The x-axis labels follow the same label ordering convention from left to right as in c). Error bars represent 68% bootstrap resampling confidence intervals. "*" means FDR corrected p-value < alpha level of 0.05, and "n.s." represents non-significant results.

Comparing the proportion of area covered between targets using different types of UFOVs (hypothesis 1)

Figure 2.8.a, top, depicts the mean PAC in the 3D search for both targets using all four UFOV types. We first consider the PAC in the 3D search for both targets using the two types of perceived UFOVs and the empirical UFOV. The difference in the mean PAC between targets in 3D with the UFOV_{PF} was -0.0615, $p = 0.0969$. For the second type of perceived UFOV, UFOV_{PP}, the difference across targets in 3D was -0.0214, $p = 0.6726$. In both instances, observers explored slightly more for the small target than the large target. However, the slight differences in the PAC for the searches of the two targets suggest that observers were exploring an equal amount of the 3D volumetric data with their perceived

UFOVs. In comparing the two perceived UFOVs (Figure 2.8.b, top), we note that the PAC ratio using the $UFOV_{P\,P}$ was significantly larger than the PAC ratio using the $UFOV_{P\,F}$ ($\bar{\Delta}ratio_{P\,P\,vs.\;P\,F;\,3D} = 0.2380, p = 0.0078$). Moreover, the PAC ratio using the $UFOV_{P\,P}$ of 1.0340 (95% bootstrap CI [0.6944, 1.4421]) produced the most consistent exploration behavior across targets for the six participants. On the other hand, when considering the target-specific empirical UFOVs, we see that the mean PAC in the 3D search for the large target was significantly greater than that for the small ($\bar{\Delta}UFOV_{E;\,3D} = 0.4407, p < 5e^{-5}$). Furthermore, when comparing the PAC ratio using the $UFOV_E$ to the PAC ratios using the three alternative UFOVs, we note that it is significantly larger than all 3 PAC ratios (Figure 2.8.b, top, and Table A.2).

How did the 2D search PAC between the two targets differ when considering the three types of UFOVs? Figure 2.8.a, bottom, confirms that what was observed in 3D held during the 2D search. Specifically, the mean difference in the PAC across the two targets with the $UFOV_{P\,F}$ was $-0.1047, p = 0.1379$. Similarly, for the second type of perceived UFOV, the $UFOV_{P\,P}$, we found no significant difference between the PAC for the two targets ($\bar{\Delta}UFOV_{P\,P;\,2D} = -0.0405, p = 0.5778$). Like in the 3D search, observers, on average, explored slightly more for the small than the large target. We also note that, like the 3D search, the mean PAC ratio utilizing the $UFOV_{P\,P}$ was marginally higher than the mean PAC ratio utilizing the $UFOV_{P\,F}$ ($\bar{\Delta}ratio_{P\,P\,vs.\;P\,F;\,3D} = 0.0675, p = 0.0499$), as shown in Figure 2.8.d, bottom. Again, the PAC ratio using the $UFOV_{P\,P}$ of 0.9502 (95% bootstrap CI [0.7806, 1.1130]) produced the most consistent exploration behavior across targets for the six participants.

Conversely, for the difference in 2D PAC between targets using the $UFOV_E$, we saw a similar result as in the 3D search. Figure 2.8.a, bottom shows that the 2D PAC with the large target $UFOV_E$ was significantly greater than that utilizing the small target $UFOV_E$ ($\bar{\Delta}UFOV_{E;\,2D} = 0.1266, p < 5e^{-5}$). As reflected in Figure 2.8.b, bottom, and Table A.2, the mean PAC ratio using the $UFOV_E$ was significantly greater than the mean PAC ratio using the standard UFOV and two perceived UFOVs. This is not surprising given the greater spatial extent of the large target $UFOV_E$ relative to the small target $UFOV_E$, exemplified for one observer in Figure 2.2.a, bottom left. With just a few fixations, one fixation in each image quadrant, one can sufficiently cover the image area with the large target empirical UFOV. This is confirmed by the fact that the mean PAC for the large target in 2D with the $UFOV_E$ was approximately 1, as shown in Figure 2.8.b, bottom.

Overall, across the 2D and 3D searches, the mean PAC using the $UFOV_{P\,P}$ was the most similar across target types. To assess the generality of our results to the chosen PC threshold of 0.82, we ran this analysis for five additional PC thresholds (0.8, 0.84, 0.86, 0.88, and 0.90). As the PC thresholds increased, the size of the $UFOV_E$, $UFOV_{P\,P,}$ and $UFOV_{P\,F}$ decreased for each participant and target combination. Figure A.5 in the appendix that regardless of the PC threshold used, the $UFOV_{P\,P}$ provided the most consistent pattern of results between the two target searches.

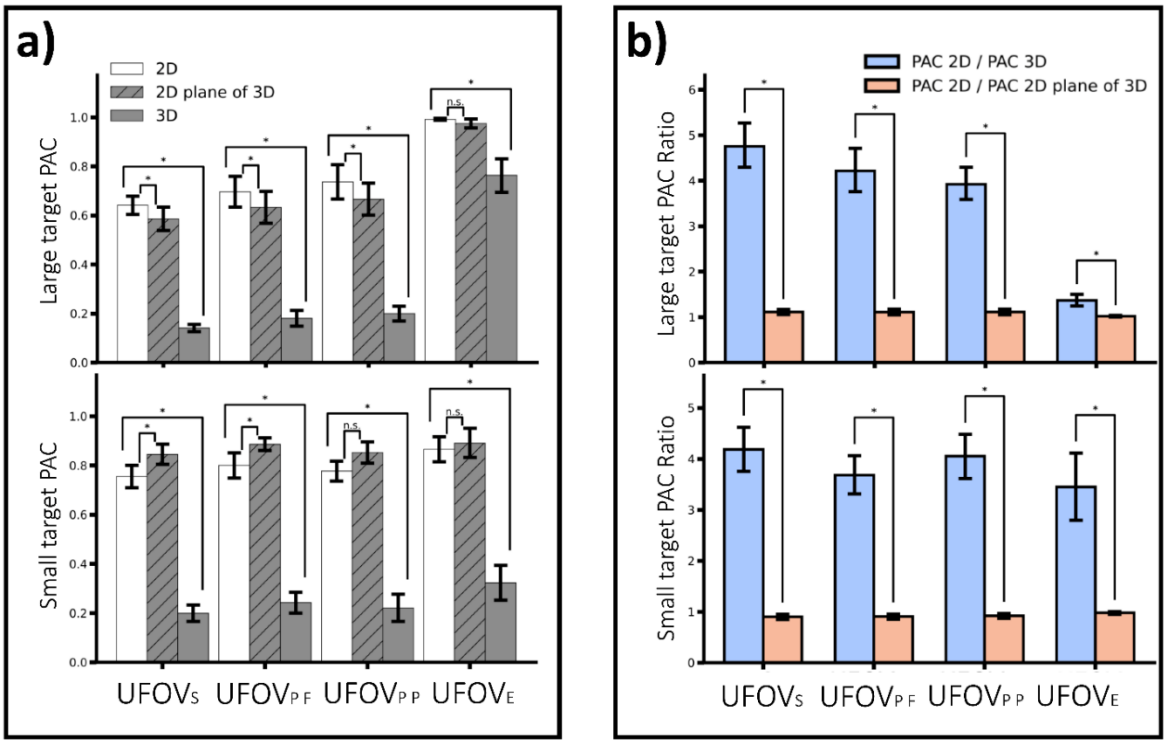

Figure 2.9. Comparing 2D and 3D search area coverage with different UFOVs by the target type. a) Mean (across observers) PAC in the 2D search (white bars), the 2D plane of 3D search (hatched gray bars), and the 3D search (gray bars). The top row depicts coverage for the large target on false negative and true negative trials, and the bottom row depicts coverage for the small target on the same types of trials. b) The ratio of PAC in 2D vs. PAC in 3D (blue bars) and PAC in 2D vs. PAC in 2D image plane of 3D (red bars). The top row corresponds to the large target search ratios, and the bottom corresponds to the small target search ratios. The x-axis represents the three types of UFOVS derived in experiment 1 and the standard UFOV for comparison. Error bars represent 68% bootstrap resampling confidence intervals. "*" means FDR corrected p-value < alpha level of 0.05, and "n.s." represents non-significant results.

Evaluating the proportion of area covered in the 2D image plane during the 3D search as a stopping criterion (hypothesis 2)

Our second hypothesis argues that the observers terminate their 3D search after sufficiently covering the 2D image plane area with their perceived UFOV. Recall that we define *sufficient coverage* as the average PAC while performing the 2D search task. Figure 2.9.a demonstrates, for each type of UFOV, the PAC in 2D, the PAC in the 2D plane of 3D, and the PAC in 3D. Based on hypothesis 1 and the consistent PAC between targets using the $UFOV_{PP}$, our following analysis focuses on the PAC in 2D versus 3D using the $UFOV_{PP}$.

We briefly comment on the PAC in 2D versus 3D using the $UFOV_E$ but refer the reader to Table A.2 for statistics concerning pairwise comparisons using the other two types of UFOVs.

For the large target (Figure 2.9.a, top), the mean PAC in 2D was marginally higher (but statistically significant) in comparison to the PAC of the 2D plane of the 3D search ($\bar{\Delta}UFOV_{P\ P;\ 2D\ vs.\ 2D\ plane} = 0.0705, p = 0.0309$). For the small target (Figure 2.9.a, bottom), the PAC in the 2D search using the $UFOV_{P\ P}$ was less but not significantly different from the PAC in the 2D plane of the 3D search ($\bar{\Delta}UFOV_{P\ P;\ 2D\ vs.\ 2D\ plane} = -0.0754, p = 0.0520$). These results suggest that when observers report the small or large target as absent, they roughly explore (7% difference on average) as much of the 2D image plane during the 3D search with their $UFOV_{P\ P}$ as they would during the 2D search for the same target.

As a point of comparison, the large target $UFOV_E$ produced a comparable PAC in the 2D search versus the 2D plane of the 3D search ($\bar{\Delta}UFOV_{E;\ 2D\ vs.\ 2D\ plane} = 0.0174, p = 0.0656$). This is not surprising given the substantial spatial extent of the large target $UFOV_E$. Interestingly, the small target $UFOV_E$ also produced a similar PAC in both the 2D search and the 2D plane of the 3D search ($\bar{\Delta}UFOV_{E;\ 2D\ vs.\ 2D\ plane} = -0.0248, p = 0.2698$). The consistent pattern in the difference in PAC across the two small-target UFOV types may be explained by the fact that observers' metacognitive estimates regarding the target's peripheral detectability (Figure 2.3, right solid line), on average, were aligned with their actual detectability of the target in their visual periphery (Figure 2.3, left solid line).

Figure 2.9.b provides another visualization suggesting the possibility search termination criterion using the 2D plane area covered. The blue bars represent the PAC ratio between the 2D search versus the 3D search, and the red bars depict the PAC ratio between the 2D

search versus the 2D plane of the 3D search. If observers quit searching in 3D after covering the 2D image plane to the same extent as in the 2D search task, we would expect the latter PAC ratio (red bars) in Figure 2.9.b to be approximately equal to 1. Indeed, both the small and large targets, regardless of the UFOV construct used, the PAC ratios cluster around 1. For example, the PAC ratio using the $UFOV_{P\ P}$ for the large target was 1.1111 (95% bootstrap CI [1.0051, 1.2424]). Similarly, the PAC ratio using the $UFOV_{P\ P}$ for the small target was 0.9186 (95% bootstrap CI [0.8353, 1.0030]).

Comparisons between the mean PAC in 2D versus the mean PAC in 3D, the white and gray bars in Figure 2.9.a, respectively, reveal that regardless of target or UFOV type, observers under-explored the 3D volumetric images with eye movements. Figure 2.9.b graphically contrasts under-exploration (blue bars) with our proposed 2D plane stopping criterion (red bars). Once again, regardless of the UFOV or target type, the former PAC ratio is substantially greater than 1. Moreover, The PAC ratio between the 2D and 3D search is significantly higher than the PAC ratio between the 2D search and the 2D plane of the 3D search.

2.5. General discussion

We sought to understand why people under-explore 3D volumetric images and how the search strategy interacted with target type. Framed another way, what evidence do observers use to quit their 3D search when they fail to find the target they were looking for? Prior work investigating the search-termination process for 2D displays has primarily taken an item-based approach (Becker et al., 2022; Lui et al., 2024; Mazor & Fleming, 2022; Shi et al., 2020). One prominent model predicting target-absent reaction times for target/distractor 2D searches suggests that the quitting signal follows a drift-diffusion process (Ratcliff,

1978) and that target-absent responses are induced once the signal surpasses an adjustable threshold modulated by estimates of local target-prevalence rates (Wolfe & Van Wert, 2010).

In this work, we take a conceptually similar approach but argue in favor of a quitting signal proportional to the image/volume area explored with eye movements—which, in theory, could be the underlying dimension on which the diffusion process occurs. This approach is particularly suited for search arrays where the distinction between target and distractor is poorly defined (i.e., no countable set of items in the display). It also can account for differences in search patterns across various targets (e.g., reaction time and number of eye movements; Figure 2.7) while emphasizing the role of extrafoveal processing on search performance (Figure 2.6). Furthermore, this approach is amenable to bridging theories of 2D and 3D searches together, of which, in the latter case, the theories are nascent and still being developed (Williams & Drew, 2019).

In short, we propose that observers keep track of the proportion of area explored with a target-specific perceived Useful Field of View. If observers do not find the target during the search, they compare this estimate to a stopping criterion and terminate their search if the estimated area explored surpasses it. The area explored with a UFOV may serve as the feature dimension along which the stopping criterion exists because this metric is positively correlated with detection performance in previous work on search in 3D volumetric images (Drew, Vo, Olwal, et al., 2013; Lago, Jonnalagadda, et al., 2021; Rubin et al., 2015). Under this working model, we tested two hypotheses. Our first hypothesis posits that when observers consider their perceived detectability of each target in their field of view, they roughly cover the same amount of image/volume area in the two separate target searches.

Our second hypothesis argues that when observers perform the 3D search and do not find the target, they quit once they have covered the 2D image plane of the 3D volume to the same extent as they would in the analogous 2D search task. Consequently, they would not explore much of the 3D area with their target-specific perceived UFOV.

To test these hypotheses, we needed a principled way of measuring each participant's metacognition regarding the spatial extent of covert attention—the perceived eccentricity at which a target can be detected for a given accuracy threshold. Experiment 1 provided a methodological framework for testing this. First, we demonstrated how people intuit target detectability in the visual periphery and how this metacognition differs from their actual detectability of a small and large target. Specifically, participants believed that the large and small targets were similarly detectable at various eccentricities (Figure 2.3, middle and right) despite them being able to detect the large target further out in the visual periphery than the small target (Figure 2.3, left). However, it should be noted that the peripheral estimation task produced an upward shift in the y-intercept of the linear fit of the large target relative to the small target by 1° in eccentricity.

The second goal of Experiment 1 was to generate target-specific empirical and perceived UFOVs for each participant. We used simple linear fits of eccentricity regressed on proportion correct to derive perceived UFOVs estimated in the fovea (task 2) or the visual periphery (task 3). We also derived empirical UFOVs for each subject by fitting a Contaminated Binormal Model (CBM) to the ROC data gathered at each eccentricity for each target. We applied a secondary fit on the CBM parameters, $\{\widehat{\mu_{cbm}}, \hat{\alpha}, \hat{\lambda}\}$, so that we could predict proportion correct for untested eccentricities (Figure 2.2.a). To normalize across the spatial extent of the empirical and perceived UFOVs, we found the eccentricity

that would predict a proportion correct value of 0.82 ($UFOV_E$). We also selected the predicted eccentricities at the same proportion correct value of 0.82 to derive the radii of the two perceived UFOVs ($UFOV_{P\,F}$, $UFOV_{P\,P}$), as shown in the subplots of Figure 2.2.b and 2.2.c, respectively.

In experiment 2, we examined the proportion area covered, or PAC, in the 2D search and 3D search for both targets using the standard UFOV with a radius of 2.5°, the empirical UFOV, and both perceived UFOVs from experiment 1. In the 2D or 3D search, the mean PAC using the target-specific empirical UFOV was significantly greater for the large target than the small one (Figure 2.8.a). However, the PAC ratio (PAC large target search / PAC small target search) was near 1 when we applied the perceived UFOVs to the observer's eye movements (Figure 2.8.b), suggesting that people may explore a similar amount of area using their perceived UFOV for both targets. The most consistent PAC between targets was found when applying the $UFOV_{P\,P}$. Interestingly, the PAC with the $UFOV_s$ and $UFOV_{P\,F}$ showed no significant difference (Table A.2), which suggests that the standard UFOV can serve as a good tool for approximating the area covered in future work investing 3D search for these types of targets.

Lastly, we found that people under-explore 3D volumetric images (Figure 2.9.b blue bars), which could be partly due to sufficient coverage of the 2D image plane with the UFOV during the 3D search (Figure 2.9.b red bars). Given the presentation interface of the 3D data to the users (i.e., scrolling through a stack of 2D images), focusing on covering the 2D image plane before quitting the search could lead to the under-exploration of the 3D volume in many ways. For example, if one scanned the upper right portion of the monitor screen at the beginning of the trial, say on slices 1-5, they may not revisit that area for some

time or at all. Moreover, they might not remember which slices they directed their gaze to in that area. As a result, there is a high probability that the upper right portion of the image stimulus on slices 6-20, for example, will only be processed by their peripheral vision. Depending on the target detectability in the visual periphery and the resultant size of the $UFOV_E$, that region of the 3D volume may go unexplored during the search.

As a brief aside, some caution should be taken in this interpretation, as prior work has shown that humans are poor at remembering where they have previously fixated during the search (Võ et al., 2016) and discerning their fixation patterns from another person's fixations patterns on the same image stimulus (Foulsham & Kingstone, 2013). However, humans may have a coarser representation of which regions of the image they have already explored with eye movements—otherwise, why would a radiologist feel confident in moving on to the next case without examining the entire medical image (Võ et al., 2016)?

It is important to consider the limitations of our study, which can limit the generalizability of our findings to real-world visual search tasks in a radiologist's office. First and foremost, we utilized trained undergraduate observers as opposed to radiologists. Undergraduate observers afforded us many trials with eye-tracking data collected over multiple weeks to procure asymptotic estimates (i.e., large N per observer) of search performance and actual versus estimated target detectability across the visual field. The observer's extensive experience with both types of targets in the simulated noise backgrounds of Experiment 1 most surely formed their search strategies in both the 2D and 3D conditions of Experiment 2, particularly when they would terminate their search. However, to translate our findings in the laboratory to real-world search scenarios, we must consider that search strategies vary with a person's expertise and experience (Nodine &

Mello-Thoms, 2010; Waite et al., 2019). For example, Drew et al. classified a sample of radiologists into two groups based on their eye movement scan paths while looking for nodules in lung CT scans: scanners and drillers. Scanners systematically foveated multiple regions in a slice before moving on to the next slice. In contrast, drillers fixated on one location in the (x, y) plane and scrolled through many slices at a time before fixating somewhere else. The radiologists with more training and experience tended to drill, whereas those with less experience tended to scan the 3D stack of images (Drew, Vo, Olwal, et al., 2013). Interestingly, drillers outperformed scanners in this experiment.

Considering our findings with undergraduate students, the discrepancy in performance between drillers and scanners may be explained by a size mismatch between the empirical and perceived UFOVs of less experienced radiologists who tend to scan the 3D image stacks. We suspect that the area of the empirical UFOV grows with a person's expertise (Lago, Sechopoulos, et al., 2020). However, we do not know how the perceived UFOV tracks with expertise. Suppose a less experienced radiologist's empirical UFOV is much smaller in area than their Perceived UFOV. In that case, it makes sense that they miss small lung nodules even after scanning the 3D image stack because the target never appears in the empirical UFOV during the search. Concurrently, they may terminate their search because they feel they explored a significant portion of the 3D image data with their perceived UFOV. For the more experienced radiologists, it could be the case that their empirical and perceived UFOVs are similar in size. Therefore, drilling allows more experienced radiologists to capture peripheral information (e.g., motion onset cues due to a small nodule flickering in and out of the peripheral field of view while scrolling through 3D), and they know that their peripheral vision is good. Because of this awareness, they know they

adequately explored most of the image data with their perceived UFOV. They are confident with terminating their search after drilling at only a subset of different locations in the (x, y) plane.

Another limitation of our study is that observers knew which target they were looking for at the beginning of each trial. This allowed us to directly measure how much of the image data they explored with both types of UFOVs. These measures of the area explored would be unattainable if they had to simultaneously search for both the large and the small targets in a single trial. However, in a clinician's office, the radiologist does not know apriori what type of lesion might be present. Microcalcifications, which we simulate as a sphere, often appear in clusters. The large target, which we model as a Gaussian blob, rarely appears symmetrical in a medical image. Spiculations and other architectural asymmetries often make the lesion appear distorted. The type of perceived UFOV a radiologist would deploy in a 3D search scenario with signal uncertainty is unclear. Would a radiologist adopt a more conservative (smaller) perceived UFOV to avoid missing potential lesions? Or would the size of the perceived UFOV depend on the patient's history and other demographic factors (i.e., risk for cancer) or a combination of both? Our framework provides the grounds for testing these hypotheses in future work.

Additional complexities about the search termination criterion arise when considering that 1) multiple lesions may be present in an actual DBT image and 2) target-prevalence influences when observers quit searching. Prior work on 2D search has demonstrated that finding one target often leads to a subsequent miss of a second target in the same image, a phenomenon known as the *satisfaction of search* (Berbaum et al., 1990; Fleck et al., 2010). The self-termination of search explains this type of cognitive error, but the hypotheses in

this paper do not directly address this circumstance. Fortunately, at screening, if one lesion is identified, then at work-up, there is a higher chance that other lesions missed on the first pass will be later identified. Regarding target prevalence rates, it is well-documented that low prevalence rates mediate when people choose to end their search. If the prevalence rate is low, people will be more prone to missing the target than if the prevalence rate is high (Ishibashi et al., 2012). Our study focused on a prevalence rate of 50% to avoid confounding our analysis with low target-prevalence effects. Nonetheless, incorporating the empirical and perceived UFOV analysis into a 3D search context with low target prevalence is an interesting future line of research.

2.6. Conclusion

3D volumetric imaging has become essential for improving performance in life-critical tasks such as early cancer detection and threat detection in airport carry-on luggage. Under-exploration of 3D image stacks has been shown to lead to search errors. Our findings suggest that the under-exploration is explained by observers' consistent strategy to use their perceived detectability of targets in the visual periphery and area explored by their eye movements in the 2D plane to terminate their search.

III. A 2D synthesized image improves the 3D search for foveated visual systems

3.1. Abstract

Current medical imaging increasingly relies on 3D volumetric data making it difficult for radiologists to thoroughly search all regions of the volume. In some applications (e.g., Digital Breast Tomosynthesis), the volumetric data is typically paired with a synthesized 2D image (2D-S) generated from the corresponding 3D volume. We investigate how this image pairing affects the search for spatially large and small signals. Observers searched for these signals in 3D volumes, 2D-S images, and while viewing both. We hypothesize that lower spatial acuity in the observers' visual periphery hinders the search for the small signals in the 3D images. However, the inclusion of the 2D-S guides eye movements to suspicious locations, improving the observer's ability to find the signals in 3D. Behavioral results show that the 2D-S, used as an adjunct to the volumetric data, improves the localization and detection of the small (but not large) signal compared to 3D alone. There is a concomitant reduction in search errors as well. To understand this process at a computational level, we implement a Foveated Search Model (FSM) that executes human eye movements and then processes points in the image with varying spatial detail based on their eccentricity from fixations. The FSM predicts human performance for both signals and captures the reduction in search errors when the 2D-S supplements the 3D search. Our experimental and modeling results delineate the utility of 2D-S in 3D search—reduce the detrimental impact of low-resolution peripheral processing by guiding attention to regions of interest, effectively reducing errors.

3.2. Introduction

A 3D volumetric medical image is typically constructed to produce an array of cross-sectional "slices" of the body anatomy (e.g., 3D breast tomosynthesis, DBT, (Chong et al., 2019; Williams & Drew, 2019)). Each slice constitutes a different plane in space, and radiologists view the slices one at a time as part of a sequence of images on a computer monitor. This design allows the radiologist to scroll back and forth through the third dimension of the reconstructed volume to visualize features of interest and segment them from the background of noise and normal anatomical structures (Aizenman et al., 2017; Georgian-Smith et al., 2019; Skaane, 2017). However, evidence suggests that radiologists do not direct their center of gaze to every region in an image or 3D volume (Krupinski, 1996; Kundel, 1975; Rubin et al., 2015). Instead, they adopt a search strategy that relies heavily on processing visual information away from points of fixation, which can be problematic when scrolling through 3D volumes (Lago et al., 2018).

Notably, peripheral vision is characterized by low spatial acuity (i.e., low sensitivity to high spatial frequencies) relative to central/foveal vision (Benson et al., 2021; Rosenholtz, 2016; Rovamo et al., 1984; Strasburger et al., 2011). Taken as a whole, the varying resolution of human vision across the visual field (~180 degrees of arc) is described as a foveated visual system. Small signals (e.g., microcalcifications), which comprise a tiny portion of the voxels in the entire 3D volume and have a radial frequency profile dominated by high spatial frequencies, are often missed in the 3D search. This can be explained as a consequence of under-exploration with eye movements and an inability to detect small signals in the visual periphery (Lago, Jonnalagadda, et al., 2021; Lago, Sechopoulos, et al., 2020). Eye-tracking studies corroborate this notion by finding a large proportion of misses

as instances where observers fail to foveate a signal (search errors) (M. P. Eckstein et al., 2018; Lago, Abbey, et al., 2021a). On the other hand, humans do not miss large mass-like signals often in 3D images because they can be more readily detected in the visual periphery (M. P. Eckstein et al., 2018; Lago et al., 2018).

Volumetric data, however, are not the only sources of visual information utilized for diagnostic purposes. Modern DBT systems make available a complimentary 2D view, either in the form of a FFDM image or a synthesized view generated from a projection of the DBT volume, which is the focus of our work. We will refer to the projection image as a 2D synthesized image or 2D-S. Currently, there is little theoretical understanding of why a complementary 2D-S image might aid detection performance and which types of signals benefit. This work aims to better understand the functional role of a projection image used in tandem with a 3D volume.

We hypothesize that the 2D-S will benefit the 3D search for small signals that are hard to detect in the visual periphery (e.g., microcalcifications). Specifically, we expect fewer search errors for the microcalcification-like signal when the 2D-S image is available. Localizing suspicious areas in the 2D-S image is relatively easy because it does not involve scrolling through a 3D stack. Our hypothesis is based on the idea that the 2D-S search guides eye movements in the corresponding 3D data, resulting in a more efficient 3D search.

To assess our hypothesis, we measured and analyzed human search performance, response times, and eye movement patterns while observers viewed images in three conditions: 2D-S, 3D, and joint presentation of 2D-S + 3D. We report results for five non-radiologist observers who searched for either signal embedded in power-law noise (Burgess et al., 2001) after task-specific training.

We then assess whether a Foveated Search Model (FSM) (Lago, Abbey, et al., 2021a) can explain human performance and errors. The model uses the measured eye movements of human observers and processes the image data at each human fixation with a set of templates that together simulate foveated vision. We reason that human search performance is heavily influenced by an interaction between eye movement exploration and peripheral detection of the signal. A model observer that captures human-foveated vision while taking in as input an observers' eye movement patterns should explain the influence of an accompanying 2D-S image on 3D search. Additionally, the model should predict how these imaging modalities interact with a signal's visibility in the visual periphery (e.g., micro calcification-like and mass-like signals). A preliminary version of the behavioral data, with fewer trials and analyses, was presented here (D. S. Klein et al., 2021).

3.3. Methods

Participants

Five graduate students from the University of California, Santa Barbara, with normal or corrected-to-normal vision, participated in this experiment. Four observers were naïve to the hypotheses of the study. One participant was the first author and was aware of the hypothesis but not the model observer predictions. The gender balance was 20% female and 80% male, ranging from 23-30 years old. All participants viewed consent forms and were treated according to the approved human subject research protocols by the University of California, Santa Barbara.

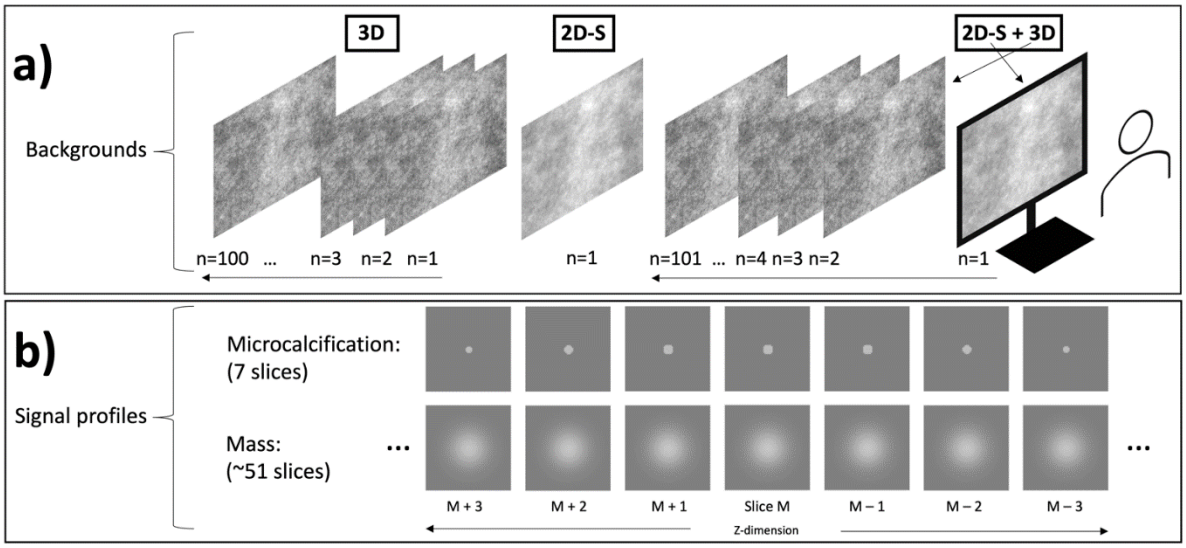

Figure 3.1. Examples of stimuli and signal profiles. Examples of stimuli used in each of the three conditions of the experiment (a). The 3D condition (left) consisted of two hundred 3D volumetric images of n=100 slices each. The 2D-S condition (middle) contained 2D synthetic views of the 3D volumes that 1) filtered out low spatial frequency information and 2) took a subsequent max operation across the third dimension. All backgrounds in the 2D-S + 3D condition (right) were composed of a 3D volumetric image, and associated 2D-S concatenated to the top of the volume (n=101, 100 slices plus the 2D-S). Observers viewed a single slice (or 2D-S) at a time on the computer monitor. Depictions of the 3D signal profiles across different slices (b). The microcalcification (top row) spanned seven slices, and the radii of the composite disks decreased away from the central slice (M), approximating a sphere. The mass, a Gaussian blob with luminance decreasing gradually away from the centroid of the signal profile, spanned approximately 51 slices.

Display and image generation

Monitor

Participants viewed stimuli on a medical-grade monitor (1,280x1,024 resolution Barco MDRC-1119 LCD monitor) 75 cm away (45 pixels per degree of visual angle or *dva*) in a darkened room (2 lux). We calibrated the monitor to have a linear contrast in luminance intensity. Specifically, 0.1 cd/m^2 and 111 cd/m^2 mapped to gray levels of 0 and 255, respectively. The background area of the monitor screen, bordering the stimulus on all sides, was set to a neutral gray level of 128.

Background

Using a pseudorandom number generator, we created all stimuli by sampling from stationary Gaussian random fields ($\mu = 128,\ \sigma = 25$). We introduced pixel-to-pixel correlations by filtering three-dimensional arrays of white noise. The filtering process simulates the idealized noise power spectrum found in mammography ($\frac{1}{f^{2.8}}$, using radial frequency indices), which characterizes the background variability (or anatomical noise) (Burgess et al., 2001; M. P. Eckstein et al., 2017). Each 3D stimulus occupied 1,024x820x100 voxels, translating to 100 "slices" (1 mm depth sampling per slice) of size 1,024x820 pixels (22.8 *dva* x 18.2 *dva*), where each pixel was converted and stored as an 8-bit integer (Figure. 3.1.a, left). A typical pixel spans 150 μm in space, making the dimensions of each slice span 19.2 cm by 15.4 cm.

Creating 2D-S

2D synthetic image generation algorithms on commercial DBT systems are proprietary. We approximated the images these algorithms produce by convolving each volume with a local filter kernel and then applying a pixel-wise max operation across the third dimension of the filtered volume, as others have done in a conceptually similar manner (H. Kim et al., 2020; S. T. Kim et al., 2014). The kernel was a 9x9x9 high-pass spherical filter, a sharpening kernel where the elements sum to one. The central elements, those less than $3.9\overline{9}$ pixels from the kernel center pixel, were set to a positive constant value of 2/251. In contrast, the elements constituting the outer shell, greater than 4 pixels away from the center but less than 5 pixels from the center pixel, were set to -1/234 so that the kernel maintained a DC frequency response of 1. The max operation was taken across the 100 slices at each (x, y) coordinate to produce a single 2D view (2D-S) of size 1,024x820 pixels (Figure 3.1.a, middle).

Signals

Next, we generated two signal profiles that approximate the geometric shape of abnormalities radiologists screen for in a routine exam. The first was a small microcalcification-like signal, a sphere spanning 7 pixels in diameter (~0.15 *dva*) and 7 slices in depth, with a uniform contrast of 0.47 (Figure. 3.1.b, top). We defined contrast as the additive luminance of the signal (microcalcification = 26.2 cd/m^2) divided by the mean luminance of the background noise (55.77 cd/m^2). The second was a larger signal, modeled as a Gaussian function in 3D ($\sigma_{x,y} = 10\ pixels = \ 0.25\ dva$, σ_z =10 *slices*) that represents a small mass lesion (FWHM in x, y was 3.5 mm). This signal had a peak contrast of 0.57 at the center and with contrast monotonically decreasing towards the edges of the signal profile (Figure 3.1.b, bottom). Both signals were linearly added to the 3D volumes to maintain their frequency-space properties. However, the frequency-space properties of the signals changed in the 2D-S images because of the filtering and max operations applied to the 3D volumes. Initially, we adopted the signal profile parameters from previous studies (Lago, Abbey, et al., 2021a; Lago et al., 2019; Lago, Jonnalagadda, et al., 2021) but chose signal contrasts to prevent ceiling effects (hit rate of 1 and false alarm rate of 0).

Psychophysics experiment: 3D, 2D-S, and 2D-S + 3D search

Apparatus

We utilized a real-time eye tracker (EyeLink Portable Duo, SR Research Inc.) to track eye movement patterns (i.e., scan paths and fixation locations) at 2,000 Hz. Participants periodically incurred calibration and validation procedures to ensure accurate recording of their eye movements. We utilized the default parameters—eye velocity and acceleration thresholds of 30 degrees/sec and 9,500 degrees/sec^2, respectively—to delineate fixations

from saccades. Scroll events, mouse clicks, and keyboard presses were all recorded at a sampling rate of 60 Hz (monitor refresh rate). The eye-tracking experiment was run through the Python programming package Psychopy (Peirce et al., 2019).

Task overview

Participants performed a Yes/No, rating and localization visual search procedure (Abbey et al., 2018; Abbey & Eckstein, 2014; Droll et al., 2009) in three different conditions: volumetric images (3D), 2D synthesized images (2D-S), and a combination of both (2D-S + 3D). They saw 600 trials (200 per condition), broken up into 30 mini-blocks of 20 trials each. Each mini-block contained 20 randomly sampled (without replacement) stimuli from a single condition.

Prior to the main experiment, participants completed two training mini-blocks per condition (6 blocks total), with feedback given at the end of each trial. Feedback included ground truth information and the presentation of the stimulus from the trial with the location of the signal marked, if applicable. Participants did not receive feedback at the end of a trial in the main experiment.

Participants had unlimited time to search for a signal (50% prevalence). We introduced signal and location uncertainty by adding the microcalcification to random locations in ¼ of the images and a mass to random locations in another ¼ of the images. Participants searched for both signals simultaneously, knowing only one would be present on any signal-present trial. Participants ended the trial by either clicking on a location where they believed the signal was present or pressing the spacebar key to indicate a signal-absent decision. Afterward, participants chose one of 3 decision options: neither present, microcalcification present, or mass present, and provided a corresponding confidence rating in their decision

using an 8-point scale. A rating of 1 indicated the highest confidence that the trial did not contain a signal, 4 indicated the lowest confidence in the signal-absent decision, and 5 and 8 indicated the lowest and highest confidence that a signal was present, respectively. In other words, participants were instructed to use ratings 5-8 only when they made a localization click and indicated that either the mass or microcalcification was present on the trial. Below, we provide details for each experimental condition.

3D search

On a given trial in the 3D search condition, participants encountered one slice of the volume at a time on the monitor screen. Participants scrolled through the volume at their leisure by toggling between 1) a mouse wheel or 2) a custom-designed scroll bar presented to the right of each image stimulus on the screen. The scrollbar allowed the participants to drag or jump across multiple slices at a time. Participants demarcated the signal's location by right-clicking on the screen to produce a single red circle overlaid on the image stimulus for visual confirmation.

2D-S search

In the 2D-S condition, participants viewed on each trial one of two hundred 2D views synthesized from the corresponding 3D volumes. Participants made localization and rating decisions like the 3D condition.

2D-S + 3D search

The 2D-S + 3D condition was like the 3D condition but with a few important caveats. First, the 2D-S image of the 3D volume was the first image participants interacted with, and the remaining one hundred images were the 3D volume slices (Figure 3.1.a right). Second, participants could left-click on the 2D-S image to produce white circles, marking suspicious

(x, y) coordinates. The circles would persist on the screen (across slices) to aid the participants in the search process. For example, the circles could serve as landmarks for saccade endpoints as participants scrolled through the 3D volume. Participants provided localization and rating decisions as described in the two previous conditions.

Figures of merit

The area under the ROC curve (AUC)

Rating data were used to construct empirical ROC curves representing the subject's (or model's) ability to discriminate abnormalities (mass or microcalcification) from images without a signal. The Area Under the Curve (AUC) was computed by integrating the empirical ROC curve (trapezoidal AUC) (Macmillan & Creelman, 2005). AUC was used to establish overall performance across the three conditions for humans and model observers.

Hit rate and false alarm rate

We also used hit rate and false alarm rate to better understand the impact of the 2D-S image on criterion-specific search performance measures. Hits were defined as signal-present trials where an observer produced a rating greater than or equal to five and selected the correct signal profile. We defined false alarms as signal-absent trials where observers produced a rating greater than or equal to five and rated one of the two signals as being present. For instance, when considering the mass signal, there were one hundred and fifty trials without a mass present in the image (fifty trials with calcifications and one hundred signal-absent trials). All of these are potential false-alarm trials.

Characterization of search

We used several parameters to quantify the time efficiency and accuracy of the search. Search time, defined as the elapsed time the stimulus was on the screen, provided us with an

overall measure of search-time efficiency. We could determine how quickly it took observers to localize the signals or quit the search under each of the three conditions. Similarly, the number of fixations, a measure positively correlated with search time, gave us a rough estimate of the amount of search space explored. Fixations were counted as changes in the center of gaze position on the (x, y) plane. We reasoned that less of the search area explored with eye movements while maintaining high localization accuracy is consistent with a time-efficient search.

We described search accuracy in terms of three metrics: search errors, recognition errors, and misses turned to hits. Search errors and recognition errors are standard metrics for assessing eye movements' role in visual search tasks (Krupinski, 2000, 2011; Kundel et al., 1978). On a given trial, a search error occurred when 1) the observer missed the signal and 2) the observer did not foveate the signal (Kundel et al., 1978). Here, we defined *foveating a signal* as a fixation location with a distance less than or equal to 2 *dva* from the center of the signal profile. A recognition error occurred when the trial outcome was a miss, but the observer made a fixation within 2 *dva* (Kundel et al., 1978). In the 3D conditions, we added a constraint to the definition of foveating a signal. Fixations needed to be within +-N slices from the center slice of the signal profile. We set N=3 for microcalcification-present trials and N=10 for the mass-present trials.

Misses turned to hits characterized whether the addition of the 2D-S image to the 3D search improved the localization accuracy of the signals relative to the 3D condition. For a given subject, this measure counted the number of 'cases' (i.e., the same stimulus ID) in which a miss was recorded in one imaging condition but correctly localized in another. The measure was reported as a proportion by dividing these counts by the total number of signal-

present trials. We stratified the data for the two signal types and the imaging condition that defined the misses (3D or 2D-S + 3D). This resulted in four conditions: microcalcification misses in the 3D search converted to hits in the 2D-S + 3D, microcalcification misses in 2D-S + 3D search converted to hits in 3D, mass misses in the 3D search converted to hits in the 2D-S + 3D, and mass misses in 2D-S + 3D search converted to hits in 3D. Of these four possible scenarios, our hypothesis would predict that a larger number of misses would be converted to hits going from 3D to 2D-S + 3D for the microcalcification signal. If there was no effect of adding the 2D-S image, then miss-to-hit proportions for both the mass and microcalcification should be approximately the same.

Statistical methods

To check overall human performance, we evaluated trapezoidal AUC using a multi-reader multi-case (MRMC) analysis that is standard for diagnostic imaging assessments. This analysis treats readers and cases as random effects and tests for differences across experimental conditions (B. Smith et al., 2022; B. J. Smith & Hillis, 2020).

For the other performance and search-efficiency measures, we utilized a non-parametric bootstrap resampling procedure. We independently resampled readers and stimuli 20,000 times for confidence intervals and significance testing on the various measures. This resampling method preserves reader and stimulus effects in the measures.

Significance was assessed for differences between the 3D and 2D-S + 3D conditions for both search time and the number of fixations without consideration for stimulus types (mass, microcalcification, and signal-absent). This allowed us to broadly focus on the impact of the 2D-S on the 3D search. For the miss to hit measure, significance was assessed both within signal type (mass and microcalcification) and across signal type for those same two imaging

conditions, which is detailed at the end of the previous section. For hit rate and false alarm rate, we considered differences across the three imaging conditions for each signal separately. For the search error rate, we again focused on the comparison between the 2D-S + 3D condition and the 3D condition but only for the microcalcification-like signal.

For each of the comparisons above, we computed a p-value as follows. For each bootstrap iteration, we computed the difference between two mean estimates—one for each condition in the comparison—to generate a distribution of difference scores. To derive a p-value for this comparison, we determined the proportion of difference scores that were less than or equal to zero. We multiplied this proportion by 2 to conduct a two-tailed test.

In addition to the 3 pairwise comparisons of the MRMC analysis, we looked at comparisons for search time and the number of fixations (2 hypotheses), miss to hit (3 hypotheses), hit rate (6 hypotheses), false alarm rate (6 hypotheses), and search error rate (1 hypothesis). In total, we evaluated 21 hypotheses of theoretical interest. We applied the Benjamini-Hochberg procedure to control the false discovery rate (FDR) at $\alpha = 0.05$ (Benjamini & Hochberg, 1995).

In the results section below, we report the FDR-corrected p-values. In Figs 4-6, observed means are plotted along with error bars that represent 95% confidence intervals for the empirical sampling distributions of our various measures of interest.

Lastly, five trials total, from three subjects, were deemed outliers and excluded from the statistical analyses because no fixations were recorded in each of those trials and we could not run the FSM on that data. A fourth subject chose not to complete the experiment and is missing 20% of all the trials—40 trials in the 2D-S condition and 40 trials in the 2D-S + 3D condition. The data can be found online here: https://dx.doi.org/10.21227/f8vk-aj29.

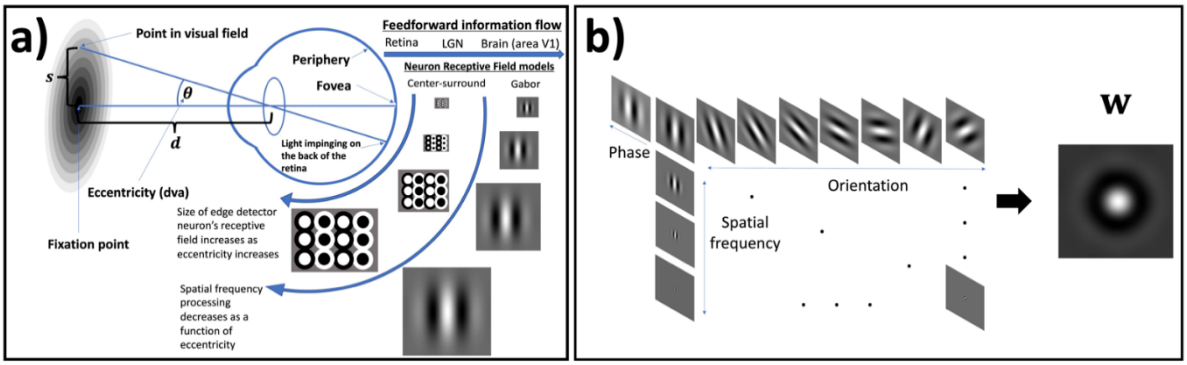

Figure 3.2. A general diagram of information flow and processing in the human visual system. Light impinges upon the retina at the back of the eye from different points in the visual field (3.2.a). For a foveated visual system, where a specific ray of light falls on the retina has downstream consequences for perceptual and cognitive processes. Signals projected onto more peripheral locations of the retina are processed with lower resolution than if they were projected onto the fovea. This is in part due to the optics of the eye. Moreover, it is due to the center-surround spatial receptive field (i.e., sensory integration area) properties of neurons in the retina and the Lateral Geniculate Nucleus (LGN) and the neurons in the visual cortex that synapse with LGN cells. Our work builds upon these general principles of signal processing in the human visual system by considering simple cell receptive fields in the visual cortex (area V1), which can function as edge detectors, typically modeled as Gabor filters. We quantify the visual periphery in terms of eccentricity, θ, in units of degrees of visual angle (*dva*). A simple geometric relation between the distance of the eye to the monitor screen, d; and the distance between the pixel of interest and the pixel being fixated at, s allows for the following mathematical relation: $\theta = \tan^{-1}\left(\frac{s}{d}\right)$. In the modeling subsections below, we relate retinal eccentricity to Gabor receptive field properties. Gabor channels are linearly combined, with optimal linear weights, to produce signal-specific model observer templates (3.2.b). On the left side of 3.2.b, we show a subset of Gabor channels used to generate a CHO template. We manipulate three parameters of the Gabor function: spatial frequency, orientation, and phase, to create a filter bank. On the right, the CHO template, tuned to the mass signal in the 2D-S background at eccentricity 9, is a linear superposition of the Gabor filters with optimal (Hotelling) weights.

Foveated Search Model (FSM)

We begin with an overview of the FSM. The central goal of the FSM is to model foveal and peripheral processing within the human visual system during the search. Signals are processed differently depending on whether the signal is foveated or not (Stewart et al., 2020). The inhomogeneous processing of signals across the visual field is due to various physiological factors: cone density in the fovea is higher than in the periphery (Curcio et al., 1990), synaptic convergence from bipolar cells to retinal ganglion cells (spatial integration) is lower in the fovea relative to the visual periphery (Purves et al., 2001), and more neurons in visual cortex per mm^2 are dedicated to processing visual information in the fovea than in

the visual periphery (cortical magnification) (Essen et al., 1984). As a result, the peripheral visual system, characterized by lower spatial resolution, limited visual processing, and lower visual sensitivity, may constrain task performance (Rosenholtz, 2016). This is particularly true in 3D images when scrolling provides less opportunity to foveate signals through eye movements (Lago, Jonnalagadda, et al., 2021).

Here, we adopt a linear model observer (Barrett et al., 1993; Burgess et al., 2001; Gifford et al., 2016; Rolland & Barrett, 1992; Sen & Gifford, 2016; Zhang et al., 2004a, 2004b) that is augmented to model known neurophysiological properties of the fovea and visual periphery. Figure 3.2.a shows a Gabor receptive field approach to modeling variable resolution in the fovea and periphery. We parametrize the visual periphery in terms of eccentricity, the angular distance (in units of *dva*) between a point in the visual field and the fixation point. Figure 3.2.a shows how eccentricity is defined.

The following sections describe the various components of the FSM model. In section "Linear observer templates", we develop the Channelized Hotelling Observer (CHO) template (Barrett et al., 1993; Rolland & Barrett, 1992; Yao & Barrett, 1992; Zhang et al., 2007), the base component of the FSM. In Section "Implementation of foveal and peripheral processing" implements foveation by changing the resolution of the linear template developed in the previous section. Specifically, the resolution of the template decreases as a function of eccentricity, reflecting the loss in spatial acuity in the visual periphery. Section "Processing image data for a given fixation" focuses on foveated processing of the image stimulus for a single fixation coordinate acquired experimentally from human observers. Section "Integrating information across fixations" describes how the FSM accumulates information at each pixel location in the 2D-S (or 3D slice) in the form of an average of log-

likelihood ratios where each ratio is generated independently for each fixation on the image. Section "Detection and localization of potential signals" describes how a final decision variable is generated along with signal-specific thresholds to make final decisions about the presence and location of signals.

Linear observer templates

The foundation of the FSM is the Channelized Hotelling Observer (CHO), inspired by the multiple-spatial-frequency channels hypothesis (Blakemore & Campbell, 1969) and first introduced into the medical imaging community in the 1980s (Myers & Barrett, 1987). This anthropomorphic model performs a detection task by reducing the image data to a set of channel responses. The final model is a linear combination of the channel responses with optimal (Hotelling) weights (Abbey & Bochud, 2000). The CHO model is typically implemented as a linear template, **w**. For search tasks, the template can be scanned over an image to localize a target.

This work utilizes Gabor channels over a set of spatial frequencies, orientations, and phases to generate a template (M. P. Eckstein & Whiting, 1995; Zhang et al., 2004a). The channels are parameterized by six spatial frequencies (32, 16, 8, 4, 2, and 1 cycles per degree of visual angle), eight orientations spaced at equal intervals between 0 and π, and even/odd phases (Watson, 1982) for a total of ninety-six channels (Figure 3.2.b, left). The channels of the CHO are arranged into a channel matrix denoted as **T**. The columns of **T** are Gabor channels lexicographically indexed such that the pixels in the 2D array representing the Gabor are remapped into a column vector. The channels are linearly independent but not orthogonal. The two-dimensional profile of the signal is also remapped to a column vector

and is denoted as **s**. The mean effect of the signal on the channels is defined by the product $\mathbf{v} = \mathbf{T}^{\mathrm{t}}\mathbf{s}$, where $\mathbf{T}^{\mathrm{t}}$ indicates the transpose of $\mathbf{T}$.

The CHO template, as shown in Figure 3.2.b, also requires the specification of an image covariance matrix representing the stochastic effects in the images (noise and anatomical variability). We denote the image covariance matrix as $\mathbf{K}_{\mathrm{g}}$. When the image is processed through the channels (i.e., multiplication by $\mathbf{T}^{\mathrm{t}}$), an image covariance matrix $\mathbf{K}_{\mathrm{g}}$ is transformed into a channel covariance matrix, $\mathbf{K}_{\mathrm{ch}} = \mathbf{T}^{\mathrm{t}}\mathbf{K}_{\mathrm{g}}\mathbf{T}$. Under the Hotelling formalism (Abbey & Bochud, 2000), the optimal linear channel weights are then given by, $\mathbf{K}_{\mathrm{ch}}^{-1}\mathbf{v}$. The resulting template is then

$$\mathbf{w} = \mathbf{T}\mathbf{K}_{\mathrm{ch}}^{-1}\mathbf{v}. \qquad (Eq.\,3.1)$$

Eq. 3.1. is adequate for 2D images. However, we also use a 3D CHO for 3D volumetric images (Lago, Abbey, et al., 2021a; Yu et al., 2017). For 3D images, we modify the CHO to include information from two slices above and below the slice under consideration. In this case, the CHO consists of five templates, $\mathbf{w}_{\mathrm{m}}$, $\mathrm{m} = \mathrm{M}-2, \ldots, \mathrm{M}+2$ (see slice index notation in Figure 3.1.b and see discussion in (Platiša et al., 2011) on ROI and the selection of adjacent slice to central slice for 3D CHO). The five CHO templates are defined as,

$$\mathbf{w}_{\mathrm{m}} = \mathbf{T}\mathbf{K}_{\mathrm{ch}}^{-1}\mathbf{v}_{\mathrm{m}}, \qquad (Eq.\,3.2)$$

Where $\mathbf{v}_{\mathrm{m}}$ is the channel matrix applied to $\mathbf{s}_{\mathrm{m}}$, one of the five central slices of the 3D signal depicted in Figure 3.1.b. The five templates in the 3D model will produce five response variables that are summed together into a single template response for a detection task. Prior work has shown that humans integrate inefficiently across time within a limited temporal window (M. Eckstein et al., 1992; M. P. Eckstein et al., 1996). We used a limited temporal integration of five slices, consistent with previous studies (Lago, Abbey, et al., 2020; Lago,

Jonnalagadda, et al., 2021). Integration of a larger number of slices increased model performance but did not vary the relative model accuracy across conditions.

The derivation so far considers detecting a signal at a known location. For search tasks with typical scanning model observers (Gifford et al., 2005), the CHO template is scanned across multiple locations. For 2D images, this is a simple 2D convolution with the image data that results in a template response at each pixel in the search region. A decision is made by comparing the max template response to a threshold. If the response is above the threshold, the model localizes the signal at that location. In 3D, a human observer sees a series of slices in the stack as they scroll through it. At each slice, the 3D model performs five 2D convolutions ranging from two slices above to two slices below the slice under consideration and sums across the five 2D responses to obtain a single 3D response at each location.

An FSM builds upon the computations a model observer uses to detect and localize signals but uses different CHO templates at different eccentricities and the next section detail its implementation.

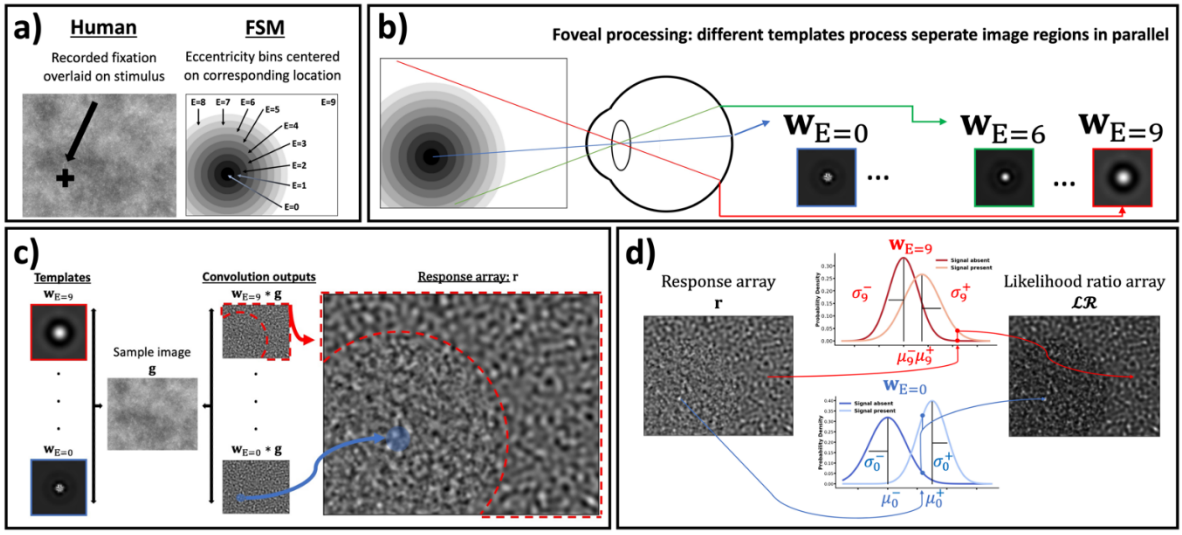

Figure 3.3. Visual processing in the Foveated Search Model for a single fixation. Foveation effects are implemented (3.3.a) using eccentricity with respect to observer fixation points. The CHO detection templates are constrained to be at lower spatial frequencies for higher eccentricity bins (3.3.b). Each template comprises bandpass filters representing visual channels, where the bands shift to lower spatial frequencies for more eccentric templates. The response array of the model (3.3.c) is generated by convolving each of the ten eccentricity templates with the image and assigning the convolutional outputs to the appropriate eccentricity bin (see text in section "Foveated Search Model (FSM)"). A likelihood ratio (3.3.d) is generated at each pixel location (l) using the ratio of signal-present and signal-absent likelihoods, as defined by response distribution properties ($\mu_E^+, \mu_E^-, \sigma_E^+, \sigma_E^-$) of each template in its respective eccentricity bin.

Implementation of foveal and peripheral processing

The FSM incorporates decreasing visual resolution as a function of eccentricity with respect to observer fixations and eye movements that reorient the fovea to regions of interest in the image. In this subsection, we describe foveal and peripheral processing for a *single* fixation.

We implement a model of peripheral effects by modifying the spatial resolution of templates according to the distance from each fixation. At each fixation, eccentricity is binned into eight non-overlapping concentric rings plus a circle centered on the fixation. Here, eccentricity ranges from zero (E=0) to nine (E=8) degrees in visual angle. We also consider the stimulus regions greater than nine *dva* (E=9). Ten eccentricity bins are sufficient to capture foveal effects while maintaining a computationally tractable model. Figure 3.3.a illustrates this binning procedure at a fixation location towards the bottom left of the image stimulus.

The loss of resolution for a template at a greater retinal eccentricity is incorporated by scaling the frequency response of the channels so that they exclude spatial frequencies at greater eccentricities. Scaling the spatial size of the Gabor channels shifts their response to lower spatial frequencies, but it maintains their 1-octave bandwidth. The scaling is implemented by a scaling constant for each eccentricity bin, E, with parameters taken from a

previous experiment that fit the model to predict d' vs. eccentricity degradation for human subjects in (Lago, Abbey, et al., 2021a). It is given by:

$$scaling(\mathrm{E}) = 1 + 0.7063\mathrm{E}^{1.6953}. \qquad (Eq.\,3.3)$$

Within each eccentricity bin, a different CHO template is used, as depicted in Figure 3.3.b. These templates are generated by modifying the channels of the channel matrix. Specifically, the wavelength, λ, of the sinusoidal component of each Gabor channel in $\mathbf{T}$ is scaled, $\lambda \rightarrow scaling(\mathrm{E}) * \lambda$. To signify the different templates used to simulate foveal and peripheral processing, we introduce an eccentricity index to indicate the scaling used on the channels, $\mathbf{T} \rightarrow \mathbf{T}_\mathrm{E}$ (E=0, …, 9). For 2D images, the resulting CHO template in each eccentricity bin is given by,

$$\mathbf{w}_\mathrm{E} = \mathbf{T}_\mathrm{E}\mathbf{K}_{\mathrm{ch,E}}^{-1}\mathbf{v}_\mathrm{E}, \qquad (Eq.\,3.4)$$

where $\mathbf{v}_\mathrm{E}$ is the product of the eccentric channel matrix and the signal, and $\mathbf{K}_{\mathrm{ch,E}}$ is the corresponding channel covariance matrix. Figure 3.3.b exemplifies three CHO templates at retinal eccentricities zero, six, and greater than nine. Similarly, for 3D images, each of the m templates is scaled for eccentricity,

$$\mathbf{w}_{\mathrm{E,m}} = \mathbf{T}_\mathrm{E}\mathbf{K}_{\mathrm{ch,E}}^{-1}\mathbf{v}_{\mathrm{E,m}}. \qquad (Eq.\,3.5)$$

Practical implementation of the FSM utilizes precomputed convolutions between the image and the CHO templates for all eccentricities. For the 2D-S images, this requires ten 2D convolutions, as shown in the left half of Figure 3.3.c. A 2D response array is generated from these ten convolution outputs by selecting the response corresponding to the appropriate eccentricity bin with respect to a given fixation, as demonstrated in the right half of Figure 3.3.c. The model produces a new response array for each fixation in the trial.

Implementation of the FSM for 3D images parallels the processing steps outlined in Figure 3.3.a-c for 2D but with a few important caveats. First, the five slice templates for each eccentricity bin defined in (*Eq. 3.5*) are concatenated together. Specifically, the 2D templates are stacked together along the third dimension with $\mathrm{m} = \mathrm{M} + 2$ at the top and $\mathrm{m} = \mathrm{M} - 2$ at the bottom of the stack of templates (recall slice index notation from Figure 3.1.b) to form a single 3D template/kernel. The 3D kernel is convolved with the volumetric data to model the varying spatio-temporal integration of information across the image data while scrolling through the slices in 3D. Second, a response array is generated only for slices where human observers make fixations. In other words, if a fixation is recorded on slice *n,* then the general process outlined in Figure 3.3.c is reproduced in the 3D search by selecting the appropriate dot products of the 3D templates and 3D image data stored on the n^{th} slice of each of the ten precomputed 3D convolution outputs.

Processing image data for a given fixation

Using eccentricity-dependent CHO templates results in response variables with different statistical properties across the range of eccentricities. To account for different statistical properties of the responses, a likelihood ratio is computed, as shown in Figure 3.3.d. This converts responses into a form that is appropriate for combining location-specific responses across different saccades (as described below in section "Integrating information across fixations"). The likelihood ratios are defined by the template response distribution parameters (means and standard deviations) of the eccentricity bin for each location in the image for a given fixation point.

Let **g** be the noisy image (2D-S or 3D), and let **r** be the template response of the FSM. The elements of **g** and **r**, g_{ι} and r_{ι} represent each pixel and the response of the FSM at that

location, respectively. For a given location, let μ_E^+, μ_E^-, σ_E^+, σ_E^- be the template response distribution parameters with respect to r_l. The eccentricity index, E, is defined for location l by the appropriate eccentricity bin for the current fixation point. As shown in Figure 3.3.d, the likelihood of a signal being present at location l is given by,

$$L_l^+ = \frac{1}{\sigma_E^+\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{r_l - \mu_E^+}{\sigma_E^+}\right)^2\right). \quad (Eq. 3.6)$$

The likelihood of the signal being absent at location l is given by,

$$L_l^- = \frac{1}{\sigma_E^-\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{r_l - \mu_E^-}{\sigma_E^-}\right)^2\right). \quad (Eq. 3.7)$$

For the 3D images, the Gaussian likelihood correctly describes the conditional distribution of linear template responses to a Gaussian-distributed image. For the 2D-S images, the likelihood calculation is approximate because of the nonlinear operation used to convert the 3D volume to a 2D synthetic image.

The model combines the likelihoods defined in (*Eq. 3.6*) and (*Eq. 3.7*) into a likelihood ratio at each location that represents the evidence for the presence or absence of the signal at location l,

$$\mathcal{LR}_l = \frac{L_l^+}{L_l^-}. \quad (Eq. 3.8)$$

Eq. 3.8 gives a likelihood ratio at each location under the assumption of a fixation position that defines the eccentricity of each location (Figure 3.3.d right).

Integrating information across fixations

We treat each observer-trial instance as a separate modeling event. As such, our implementation of the FSM utilizes an individual human observer's scan path (i.e., the time-sorted sequence of fixations and scrolls) to model the response of that subject in each trial.

The following approach for integrating information across fixations is applied to the 2D-S images and 3D volumes on slices with fixations.

Since there are generally multiple fixations within a given trial, a single location in the image may be incorporated into the modeling by appearing in multiple eccentricity bins. This means the observer may process the same region of an image with varying spatial resolution (CHO templates) across the sequence of fixations. The FSM accounts for this by accumulating likelihood ratios across fixations for each pixel location, l separately.

The likelihood ratio in *Eq. 3.8* depends on a particular fixation position. For multiple fixations, k=1, …, K, the model accumulates log-likelihood ratios and then divides by the number of fixations to produce a test statistic,

$$\lambda_l = \frac{1}{K}\sum_{k=1}^{K} \log(\mathcal{LR}_{l,k}). \qquad (Eq. 3.9)$$

Note that the number K will vary across trials (and slices in 3D) for each observer. Therefore, λ_l—which is conceptualized as the average response of the FSM at the end of the trial for a given location—will combine potentially different CHO template responses depending on the eccentricity of location l with respect to the various K fixation points in the image.

Detection and localization of potential signals

After λ_l has been computed for each location l in the 2D-S image or 3D volume, the FSM generates a signal-specific decision variable by taking the maximum across all possible locations,

$$\lambda^{\mathrm{FSM}} = \max_{l}(\lambda_l)\,. \qquad (Eq.\ 3.10)$$

So far, the development of the FSM, from the definition of the CHO templates to the accumulation of log-likelihood ratios, has assumed a specific signal profile. However, there are two possible signals in the search task and three possible decisions: mass-present, microcalcification-present, and signal-absent. To account for signal uncertainty, the FSM produces two decision variables per trial, $\lambda_{\text{Mass}}^{\text{FSM}}$ and $\lambda_{\text{Micro}}^{\text{FSM}}$, where the subscript indicates the signal profile being modeled.

The FSM converts these decision variables into a final decision by comparing $\lambda_{\text{Mass}}^{\text{FSM}}$ and $\lambda_{\text{Micro}}^{\text{FSM}}$ to signal-specific thresholds. If both response variables exceed their respective thresholds, the model selects the signal associated with the larger response. If both responses are less than their respective thresholds, the model responds "absent." In the two cases where one response is greater than its respective threshold and the other is not, the model selects the signal associated with the response variable greater than its threshold.

The FSM assumes that human observers maintain multiple internal thresholds for detection. The FSM instantiates this notion by computing sets of thresholds independently for each participant. There are six thresholds per participant, one for each imaging condition and signal type combination. For determining a threshold, we first compute the ROC curve for the model by treating $\lambda_{\text{Mass}}^{\text{FSM}}$ (or $\lambda_{\text{Micro}}^{\text{FSM}}$) as ratings. As with the human subjects when considering the criterion-specific measures of performance, there are 50 signal-present trials and 150 signal-absent trials per condition. Next, we find the operating point, (hit rate, false alarm rate), on the model ROC curve closest to the human operating point in the ROC space. The detection threshold producing that operating point becomes the threshold for a given signal and imaging condition combination.

We explored two additional common choices for model thresholds, one that maximizes proportion correct and another that generates a match between human and model false alarm rates. The three methods to select decision thresholds resulted in the same relative model performance across conditions.

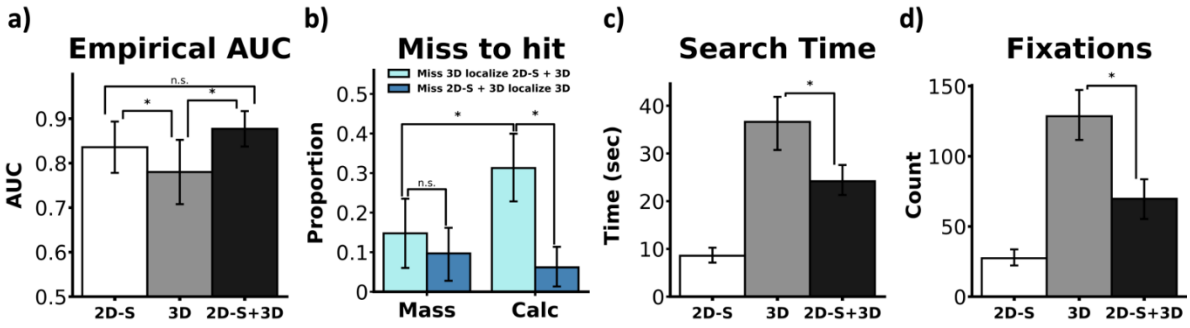

Figure 3.4. Evaluating search performance and efficiency for three imaging modalities. Area Under the ROC Curve considers the participants' confidence ratings in the signal-present trials and the confidence ratings in the absent trials (3.4.a). It provides an overall measure of the effectiveness of the 2D-S during the 3D search. The misses turned to hits were computed using the trials from the 2D-S + 3D and 3D conditions (3.4.b). We collapsed search time (3.4.c) and the number of fixations (3.4.d) across signal types in each condition. Error bars represent 95% confidence intervals (CI) that consider reader and case variability. CIs in 3.4.a are estimated from a Linear Mixed Effects MRMC model whereas CIs in 3.4.b-3.4.d are generated from empirical bootstrap resampled distributions. * = FDR-adjusted p value < threshold at $\alpha = 0.05$, n.s. = non-significant.

3.4. Results

2D-S serves a functional role in 3D search

Our main interest is the impact of the 2D-S image in 3D search. Specifically, we determine whether the 2D-S, serving as an adjunct to the 3D volumetric image, improves performance, and if so, how. The MRMC analysis found a significant effect of the reading condition ($F(2, 12.80642) = 11.92141$, $p = 0.00119$) with performance in 3D imaging alone significantly less than 2D-S + 3D ($\Delta AUC_{\text{2D-S + 3D vs. 3D}} = 0.09702$, $p = 0.00097$) and 2D-S ($\Delta AUC_{\text{2D-S vs. 3D}} = 0.05568$, $p = 0.03604$). The 2D-S + 3D condition improved performance

over 2D-S alone but was not significant ($\Delta AUC_{2D\text{-}S + 3D \text{ vs. } 2D\text{-}S} = 0.04134, p = 0.11245$). As depicted in Figure 3.4.a, these findings are congruent with the general pattern of results from our bootstrap resampling procedure used in subsequent analyses.

Figure 3.4.b shows the average miss-to-hit proportions across subjects. Adding the 2D-S to the 3D search increases the miss-to-hit proportion significantly for the microcalcification-like signal (Δproportion = 0.24995, p = 0.00126). Furthermore, 2D-S + 3D converts a significantly greater proportion of misses to hits for the microcalcification signal than for the mass signal (Δproportion = 0.16875, p = 0.02074). For the mass signal, there is no significant increase (Δproportion = 0.05045, p = 0.45791) for the miss-to-hit proportion going from 3D to 2D-S + 3D.

The average time to complete a trial (Figure 3.4.c) and the average number of fixations per trial (Figure 3.4.d) are displayed for the three imaging conditions. Not surprisingly, the search times in the 2D-S condition are much faster than in either 3D search condition. In the 2D-S + 3D condition, both the average time to complete a trial and the average number of fixations were significantly lower relative to the 3D condition (ΔTime = 12.16435 seconds, $p < 5e^{-5}$; ΔFixations = 57.89218, $p < 5e^{-5}$). Taken together, the 2D-S image used in conjunction with the 3D volumetric image makes the search more time-efficient, reduces misses, and improves accuracy.

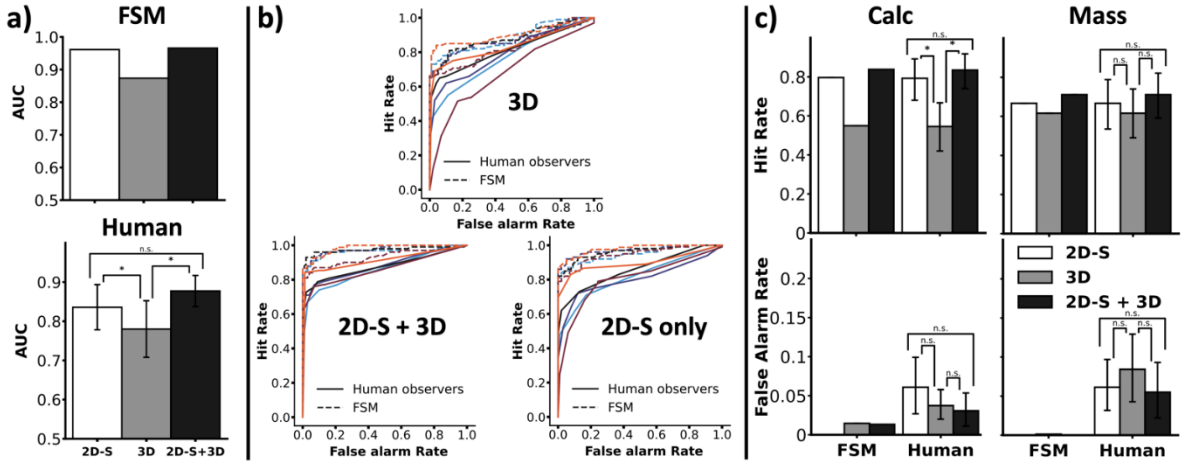

Figure 3.5. FSM vs. human performance across imaging conditions (a, b) or imaging condition and signals (c). FSM AUC compared to human observer AUC (3.5.a). The top figure denotes the mean FSM empirical AUC for each condition (i.e., treating calc-present and mass-present trials as "signal-present" trials). The bottom figure depicts the human AUC results from Figure 3.4.a, reproduced here for visual comparison to FSM performance. FSM vs. human observer ROC curvers across imaging conditions (3.5.b). The curves in each of the three plots are computed by subject ID (color) and observer type (linestyle: solid-human, or dashed-model). The curves are used to generate the AUC results in 3.5.a. FSM vs. human comparison on criterion-specific measurses of performance for mass and microcalcification-like signals in each of the three conditions (3.5.c). The top left panel illustrates the average hit rate of the FSM (left three columns) and human observers (right three columns) for the microcalcification signal. Each colored column represents a different imaging condition. The top right panel represents the average hit rate for the mass signal also stratified by the observer (left column cluster-FSM, right-human) and condition (column color). The bottom two panels are organized graphically in the same manner as the top row, but the dependent variable is the average false alarm rate. Error bars in Figure 3.5.a bottom are 95% CIs estimated from the MRMC analysis and error bars in Figure 3.5.c represent bootstrap distribution 95% CIs that consider reader and case variability. * = FDR-adjusted p value < threshold at $\alpha = 0.05$, n.s. = non-significant.

Comparing human vs. model performance

The second aim of this experiment is to show that foveal effects, as implemented in the FSM, can predict human performance in these search tasks. Figure 3.5 demonstrates that the FSM produces similar performance trends for the two signals across the three imaging conditions compared to the human observers. First, Figure 3.5.a replots the MRMC analysis from Figure 3.4.a underneath FSM AUC performance across the three imaging conditions (Figure 3.5.a, top panel). The mean differences in the FSM's AUC for each combination of conditions are similar to the performance trends of human observers despite the model performing better overall. For example, the mean AUC for the FSM in the 3D imaging alone

condition is less than 2D-S + 3D ($\Delta AUC_{2D\text{-}S + 3D \text{ vs. } 3D} = 0.09202$) and the 2D-S condition ($\Delta AUC_{2D\text{-}S \text{ vs. } 3D} = 0.08759$). These results are on par with mean differences in AUC across conditions for the human subjects. However, the FSM's mean difference in AUC between the 2D-S + 3D condition and the 2D-S alone ($\Delta AUC_{2D\text{-}S + 3D \text{ vs. } 2D\text{-}S} = 0.00443$) diverged substantially from the average difference in human AUC across these two conditions, which was 0.04134.

Figure 3.5.b highlights the absolute differences between human and FSM performance across the three conditions. By visually inspecting the ROC curves in Figure 3.5.b, we see that curves generated from the human observer rating data (solid lines) tend towards the chance line ($y = x$) more relative to the ROCs generated from the FSM response variables defined in *Eq. 3.10*. This is true for all imaging conditions. Despite this, the FSM captures individual differences in search performance, which is demonstrated by the color coding of the ROC curves. Each color represents a different participant. The solid lines are the observer ROCs and the dotted line with the same color is the corresponding FSM ROC using that observer's fixations.

Next, we looked at hit rate and false alarm rate. The top row of Figure 3.5.c compares the hit rate of the humans to the FSM for each signal in all three conditions. For the microcalcification signal (Figure 3.5.c, top left panel), human observers maintain significantly higher hit rates in the 2D-S + 3D and 2D-S condition relative to the 3D condition ($\Delta HR_{2D\text{-}S + 3D \text{ vs. } 3D, \text{ calc}} = 0.28889$, $p < 5e^{-5}$, $\Delta HR_{2D\text{-}S \text{ vs. } 3D, \text{ calc}} = 0.24715$, $p = 0.00105$). The 2D-S + 3D condition HR is not significantly higher, for humans, relative to the 2D-S condition ($\Delta HR_{2D\text{-}S + 3D \text{ vs. } 2D\text{-}S, \text{ calc}} = 0.04175$, $p = 0.42136$). The FSM produces the same pattern as humans for differences in mean hit rate across the three conditions ($\Delta HR_{2D\text{-}S}$

$_{\text{+ 3D vs. 3D, calc}}$ = 0.28827, $\Delta HR_{\text{2D-S vs. 3D, calc}}$ = 0.24706, $\Delta HR_{\text{2D-S + 3D vs. 2D-S, calc}}$ = 0.04120). For the mass signal, we do not see any significant differences across the three conditions for the humans ($\Delta HR_{\text{2D-S + 3D vs. 3D, mass}}$ = 0.09524, p = 0.14280, $\Delta HR_{\text{2D-S vs. 3D, mass}}$ = 0.05079, p = 0.41335, $\Delta HR_{\text{2D-S + 3D vs. 2D-S, mass}}$ = 0.04445, p = 0.34243). The FSM mirrors smaller differences in human mean hit rate across the three conditions for masses ($\Delta HR_{\text{2D-S + 3D vs. 3D, mass}}$ = 0.09524, $\Delta HR_{\text{2D-S vs. 3D, mass}}$ = 0.05079, $\Delta HR_{\text{2D-S + 3D vs. 2D-S, mass}}$ = 0.04445).

We find no significant differences in the human observer microcalcification false alarm rates across the three conditions. The pattern of non-significant differences holds for the mass signal as well (Figure 3.5.c, bottom row). There are qualitative differences between human observers and the FSM in average false alarm rates for both signals in all three conditions. Human observers tend to make few false alarms for both signals whereas the false alarm rate for the FSM is either close to or at zero.

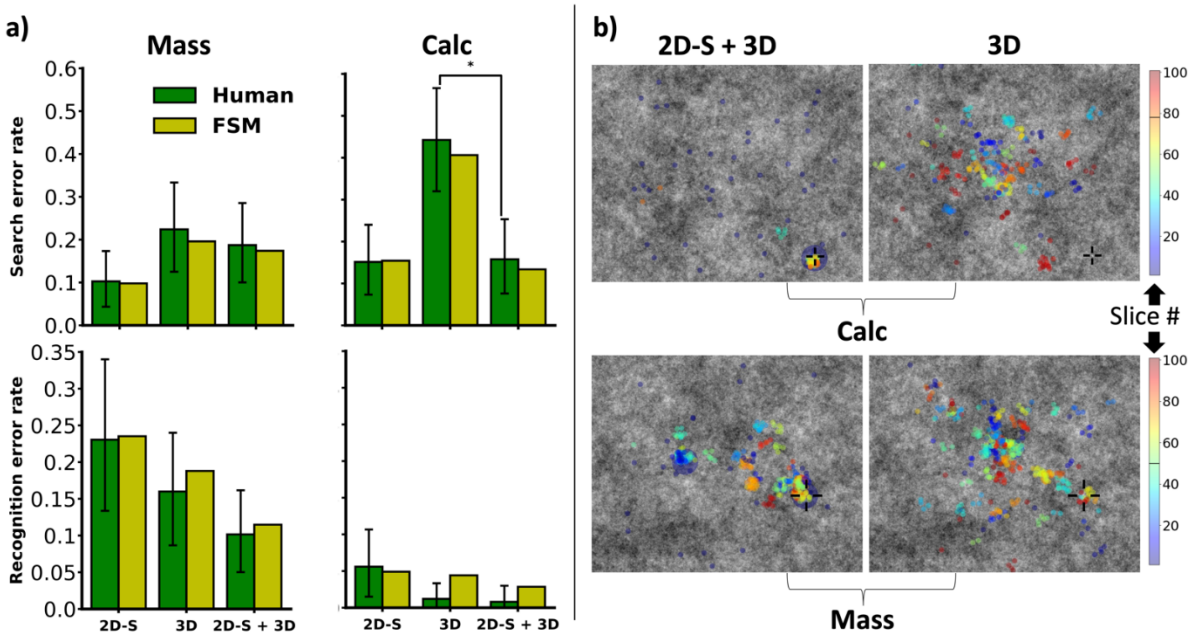

Figure 3.6. 2D-S guides eye movements to reduce misses. Search and recognition error rates in all three conditions for humans and the FSM (3.6.a). The top row shows mass (left), and microcalcification (right) search error rates across all three conditions. Yellow bars correspond to model error rates, and green bars correspond to human error rates. Bottom row: same as top row but for recognition error rates. Error bars represent bootstrap distribution 95% CIs that consider reader and case variability. * = FDR-adjusted p value

< threshold at $\alpha = 0.05$, n.s. = non-significant. Two examples of observers missing the signals in 3D search but localizing them with the 2D-S available (3.6.b). In the top row, one participant searches through the same 3D volume with the 2D-S (left column) or without the 2D-S image (right column). They localize the microcalcification in the 2D-S image and scroll to the signal in the 3D volume. In the bottom row, another observer localizes the mass in the 2D-S + 3D condition but misses it in the 3D condition. Small circles indicate observer fixation locations in the (x, y) plane throughout the trial. The colors of the circles denote the slice number in the third dimension. Blue circles correspond to fixations closer to the top of the image stack where participants begin their search, and red circles correspond to fixations on slices towards the bottom of the image stack. Larger blue circles in the left column denote clicks made by the observer on the 2D-S image to mark suspicious locations on the (x, y) plane. The black horizontal lines on the color bars indicate the central slice number of the 3D signals. The black fiduciary marks surrounding both signal profiles emphasize the signals' location for display purposes (not shown to observers). All four images contain the center slices of the respective signal profiles in 3D.

2D-S guides eye movements and reduces search errors in 3D search

To assess the role of foveation in all three search conditions, we compare the human search and recognition errors to the FSM search and recognition errors because the model explicitly incorporates foveation into the decision-making process. Figure 3.6.a top row demonstrates that the model observer's search error rates (SER) were consistent with the SER of the human observers for both signals in all three conditions. Of particular interest to the main hypothesis of this study is the difference in SER for the microcalcification-like signal in the 3D condition versus the 2D-S + 3D condition for human observers. The $\Delta SER_{\text{2D-S + 3D vs. 3D, calc}} = 0.28481$, $p < 5e^{-5}$, suggests that the 2D-S facilitated 3D search by reducing the number of search errors for the small signal. The 2D-S could have guided the center of gaze to probable signal locations in the (x, y) plane that could later be scrutinized more thoroughly in the 3D volume.

Figure 3.6.b exemplifies the moderating effect of the 2D-S on the 3D search for both signals. The observers miss the signals in the 3D condition and never fixate on them. In the 2D-S + 3D condition, the observers fixate the signals in the 2D-S and localize them in the 3D volume afterward.

The FSM recognition error rate (Figure 3.6.a bottom row) for the mass signal was similar to humans. The FSM generated higher recognition error rates for the microcalcification-like signal than humans in the 3D and 2D-S + 3D conditions.

3.5. Discussion

We are interested in understanding the perceptual impact of the 2D-S serving as an adjunct to the 3D volumetric image in the search task. Specifically, we demonstrate that accuracy improves, and the search process becomes more time-efficient when we add the 2D-S image to the 3D search. The AUC is highest when observers utilize the 2D-S image in the 3D search (Figure 3.4.a). However, the 2D-S markedly improves the 3D search for the small microcalcification signal while modestly improving the detection of the mass signal, as noted in Figure 3.4.b. Additionally, when the observers scan the 2D-S image before starting the 3D search, the speed to complete the trial and the number of fixations decreases significantly (Figure 3.4.c and Figure 3.4.d, respectively). The improvements in performance and reduction in search time and the number of fixations suggest that observers utilize the 2D-S extensively. We attribute the observed higher AUC in the 2D-S condition relative to the 3D condition (Figure 3.4.a) to the decrease in microcalcification search errors for the 2D-S condition (Figure 3.6.a top right).

Our second goal is to demonstrate that a model simulating human foveation can account for the human-observer performance in all three conditions for both signals. First, comparing accuracy across conditions, our results show that the difference in average AUC of the human observers across the three conditions matches the pattern of differences in AUC scores for the FSM across the conditions (Figure 3.5.a). Second, the FSM, taking the human-observer fixation positions as input, matches human observer hit rates across the

three conditions for both the microcalcification and the mass signals. However, the false alarm rates are either extremely low or not present for the FSM but present for humans (mass). For the microcalcifications, the FSM shows false alarm rates that do not follow those of humans across the three conditions (microcalcification). Additionally, the FSM recognition errors are higher than the human observers in the 3D search condition for the microcalcification, suggesting that there might be a systematic difference between the human and model 3D perceptual templates (Abbey et al., 2018; Abbey & Eckstein, 2007, 2014).

Taking the model predictions one step further, we glean new insights about how the 2D-S image complements the foveated nature of the human visual system to confer a benefit in the 3D search task. Specifically, the model explains the accuracy benefits of the 2D-S when accompanying the 3D images. It also explains the interaction of the 2D-S benefits and a signal's visibility in the visual periphery. We observe fewer search errors for both humans and the FSM in the 2D-S + 3D condition relative to the 3D condition, with a substantially reduced search error rate for the microcalcification signal (Figure 3.6.a, top right). This finding aligns with our hypothesis that the under-exploration of 3D volumetric images and the low detectability of small signals in the visual periphery leads to increased search errors in the 3D search. If foveation is not mediating the discrepancy in error rates between the 2D-S + 3D condition and the 3D condition, we would expect the humans to make fewer search errors in the 3D condition, which would contrast with the FSM search error rate.

This study has multiple limitations that need to be considered. For example, the 3D backgrounds we generated, $1/f^{2.8}$ filtered white noise, share a common NPS with mammographic images (Abbey & Barrett, 2001), but diverge significantly in appearance

from more realistic simulations (Bochud et al., 1999; Castella et al., 2008) and digital breast phantoms (Bakic et al., 2018; Barufaldi et al., 2018). For masses embedded in more realistic backgrounds containing normal anatomy, 3D volumetric images provide radiologists with a reconstructed image that allows them to segment the mass from anatomical noise much better than when viewing a 2D reconstructed image. Therefore, our experiment might underestimate the effect size in reducing recognition errors between 3D reconstructions versus 2D images. Similarly, our algorithm used to generate the 2D-S images filters out low spatial frequency image information, thus increasing the SNR for the microcalcification signal. Not all commercially available 2D-S algorithms may do this, leading to potential differences in how the 2D-S facilitates 3D search in a clinical setting (Nelson et al., 2016).

Beyond the simplified stimuli in this study, we utilized trained student observers instead of actual radiologists. Trained observers afforded us many trials with eye-tracking data but at the expense of limiting our ability to generalize our findings to clinical settings.

However, our previous studies have shown that some of the bottlenecks in search obtained with synthetic images and trained student observers generalize to more realistic phantoms and radiologists (M. P. Eckstein et al., 2017; Lago et al., 2017, 2018). In particular, under-exploration of the 3D images and low visibility of signals in the visual periphery led to search errors for radiologists and trained observers (M. P. Eckstein et al., 2017; Lago et al., 2017, 2018). In addition, our findings agree with studies with radiologists showing the benefits of adding a 2D image to the 3D volumetric data (Skaane et al., 2014; Zuley et al., 2014). Taken together, our study suggests the main contribution of the 2D images is to guide the search for small signals in the 3D volume and might apply well to clinical settings where radiologists routinely scrutinize 3D images such as DBT data.

Lastly, we note some of the limitations of the FSM. First, the FSM depends on the eccentricity scaling estimated from previous work which utilized similar signals in 1/f noise images (Lago, Abbey, et al., 2021a). The eccentricity scaling of the model may not generalize to other anatomical backgrounds and signals. The second limitation is the human fixation points, which serve as an input to the model might not be always accessible. Other versions of the model include an eye movement algorithm that generates a sequence of exploratory eye movements as described recently in (Lago, Abbey, et al., 2021a; W. Zhou & Eckstein, 2022).

3.6. Conclusion

A complementary 2D synthesized image can reduce errors in the 3D search for small signals that are otherwise missed due to under-exploration of the 3D volume and low signal detectability in the visual periphery. The 2D- synthesized image serves to guide eye movements during the 3D search. Predicting such effects requires a model observer that incorporates properties of the human visual system when processing information across the entire visual field (foveation). Together, our findings show how visual psychophysics, eye tracking, and model observers incorporating foveation can explain human search performance bottlenecks and help guide the assessment of 3D medical image quality.

IV. Greater benefits of deep learning-based computer-aided detection systems for finding small signals in 3D

4.1. Abstract

Radiologists are tasked with visually scrutinizing large amounts of data produced by 3D volumetric imaging modalities. Small signals can go unnoticed during the 3D search because they are hard to detect in the visual periphery. Recent advances in machine learning and computer vision have led to effective computer-aided detection (CADe) support systems with the potential to mitigate perceptual errors. Sixteen non-expert observers searched through digital breast tomosynthesis (DBT) phantoms and single cross-sectional slices of the DBT phantoms. The 3D/2D searches occurred with and without a convolutional neural network (CNN)-based CADe support system. The model provided observers with bounding boxes superimposed on the image stimuli while they looked for a small microcalcification signal and a large mass signal. Eye gaze positions were recorded and correlated with changes in the area under the ROC curve (AUC). The CNN-CADe improved the 3D search for the small microcalcification signal ($\Delta\, AUC = 0.098, p = 0.0002$) and the 2D search for

the large mass signal ($\Delta\, AUC = 0.076, p = 0.002$). The CNN-CADe benefit in 3D for the small signal was markedly greater than in 2D ($\Delta\Delta\, AUC = 0.066, p = 0.035$). Analysis of individual differences suggests that those who explored the least with eye movements benefited the most from the CNN-CADe ($r = -0.528, p = 0.036$). However, for the large signal, the 2D benefit was not significantly greater than the 3D benefit ($\Delta\Delta\, AUC = 0.033$, $p = 0.133$). The CNN-CADe brings unique performance benefits to the 3D (vs. 2D) search of small signals by reducing errors caused by the under-exploration of the volumetric data.

4.2. Introduction

Digital breast tomosynthesis (DBT) is becoming the standard imaging modality for early cancer screening within the United States (Health, 2023). DBT affords a quasi-3D rendering of the patient's anatomy that reduces tissue superposition and signal occlusion inherent in the 2D planar views generated from digital mammography image reconstruction algorithms (Sechopoulos, 2013). Radiologists interpret DBT images by freely scrolling back and forth through cross-sectional slices of the volumetric data—displayed one at a time on a computer monitor—to visually segment masses, microcalcifications, and architectural distortions from surrounding parenchyma (Helvie, 2010).

3D volumetric images, however, pose new challenges to the radiological decision-making process (Williams & Drew, 2019) because of the increased data requiring visual scrutiny. It would be prohibitively time-consuming to scan exhaustively, with eye movements, each cross-sectional slice in the stack of images before terminating one's search. Therefore, radiologists must adopt new search strategies to perform 3D visual searches (Aizenman et al., 2017; Drew, Vo, Olwal, et al., 2013; M. P. Eckstein et al., 2018). For example, a recent eye-tracking study demonstrated that radiologists and trained human

observers rely on peripheral vision when scrolling through 3D volumetric images. Due to under-exploring the 3D image stack with eye movements, trained observers and radiologists miss small signals that are hard to detect in the visual periphery (Ba et al., 2020; Lago, Jonnalagadda, et al., 2021). Specifically, under-exploration leads to search errors of small signals, a miss that occurs because the observer failed to direct their center of gaze to the signal's location (Kundel, 1989; Kundel et al., 1978).

Recent advances in deep learning-based computer-aided detection (CADe) algorithms provide a promising avenue for mitigating search errors in 3D volumetric images. First, unlike human observers, Convolutional Neural Network (CNN)-based CADe systems are not constrained by attentional bottlenecks that are a consequence of foveated vision—high spatial acuity in the fovea and low spatial acuity in the peripheral visual field (Stewart et al., 2020). The convolution kernels in a CNN can process each voxel in a large 3D volumetric image in parallel while simultaneously filtering for both high and low spatial frequency information (Yamashita et al., 2018). Second, CNN-based CAD algorithms—models that can perform both classification (CADx) and detection simultaneously—have obtained non-inferior performance relative to expert radiologists (Kooi et al., 2017; Rodríguez-Ruiz et al., 2019). Thus, these artificial intelligence-based support systems can work in parallel with an attending radiologist as a "co-pilot" to enhance and augment their workflow (Conant et al., 2019; Yang et al., 2022).

To date, no systematic vision science investigation delineates how a CNN-CADe algorithm benefits visual search in 2D versus 3D imaging modalities. For instance, does the CNN-CADe induce different performance benefits for 2D and 3D imaging modalities, and do these benefits depend on whether the signal is spatially large or small? Moreover, what

types of errors does the CADe system mitigate in 2D? Are they the same types of errors as in 3D? Do individuals who under-explore the image/volume with eye movements benefit the most from the additional information provided by the CNN-CADe adjunct?

To answer these questions, we conduct an eye-tracking study to evaluate the utility of a CNN-CADe support system on human detection performance. The model produces bounding boxes on suspicious locations made viewable to naive (trained) observers while they perform a visual search task for simulated cancers in DBT phantoms (50% prevalence rate). Specifically, trained observers search with (and without) the CNN-CADe for a small microcalcification-like signal and a large mass-like signal embedded in 3D breast phantoms (3D search) and single slices of the phantoms (2D search).

We hypothesize that the CNN-CADe will guide an observer's eye movements to suspicious locations in the 3D volumetric image that would have otherwise been missed without it. We predict a more considerable reduction in microcalcification search errors in 3D than in 2D because it is relatively easy to explore most regions of a 2D image with eye movements in a time-efficient manner. For the large mass-like signal, we hypothesize that the search with the CNN-CADe will result in a less pronounced reduction in search errors in 3D because the mass is more detectable in the visual periphery than the small microcalcification-like signal. However, the mass-like signal is more difficult to recognize in 2D than 3D. Thus, we predict that the CNN-CADe will mitigate 2D recognition errors—misses that occur even after fixating the signal (Krupinski, 1996). Finally, we hypothesize that observers with the highest degree of under-exploration of 3D images will benefit the most from the CNN-CADe adjunct when searching for a small microcalcification-like signal. To quantify an observer’s personalized, effective exploration of 3D volumes, we

combine their eye movement scan path data with an estimated Useful Field of View acquired from a separate task that measures an observer's peripheral detectability for each signal.

4.3. Methods

Participants

Sixteen undergraduate students (62.5% female, age range 18-22) from the University of California, Santa Barbara, participated in this experiment for course credit. All participants provided informed written consent and were treated according to human subject research protocols approved by the University of California, Santa Barbara (protocol # - 12-23-0301). Participants maintained normal or corrected-to-normal vision throughout the duration of the experiment.

Apparatus

Display monitor

Participants interacted with stimuli on a medical grade grayscale DICOM monitor (Brand-Barco, type-MDNG-6121; 24 Hz refresh rate; 5.8 MP or 2096x2800 pixel resolution or 325x430 mm screen size) at a viewing distance of 750 mm in a darkened room (ambient luminance = 2 lux). 45 pixels on the monitor screen subtended 1 degree of visual angle (dva). We calibrated the monitor with a Barco LCD sensor (42630), and it passed a MediCal QAWeb DICOM GSDF compliance test with a maximum error of 7%.

Eye-tracker

While participants engaged with the task, an eye tracker (SR Research Eyelink Desktop Mount) monitored their gaze position at 2000 Hz. Participants encountered a calibration and validation procedure at the beginning of each session and could recalibrate between trials if

necessary. Each procedure used a nine-point grid, and successful calibration was met if the average validation error across the 9 grid points was less than 1 dva and the max error was less than 1.5 dva. Fixations and saccades were analyzed offline using the standard velocity and acceleration thresholds of 30 deg/s and 9,500 deg/s^2, respectively.

Experiment control

The experiment used the Python package PsychoPy (Peirce et al., 2019). Events such as mouse scrolls and clicks were sampled at the monitor refresh rate of 24 Hz but synced to a wall clock via the ioHub event monitoring module in Psychopy to facilitate co-registration in the timing of these events with saccade and fixation data acquired from the eye tracker.

Stimuli

Phantoms

Participants viewed anthropomorphic DBT phantoms that simulate the spatial arrangement of anatomical tissues (skin, Cooper's ligaments, adipose, and glandular) and lesions (microcalcifications and masses). The phantoms were generated with the OpenVCT virtual breast imaging tool from the University of Pennsylvania (Bakic et al., 2018; Pokrajac et al., 2012; Predrag R. Bakic, 2017) using clinical acquisition geometry and clinical automatic exposure control settings (Selenia Dimensions, Hologic, Marlborough, MA). The 700 ml simulated phantoms were compressed in the mediolateral direction at 6.33 mm thickness with glandular tissue prevalence of 15%-25%. The spatial reconstruction parameters were set to 100 μm in-plane resolution and 1 mm depth sampling (Briona Standard; Real Time Tomography, LLC, Villanova, PA), producing a 3D voxel array of size 2048x1792x64. Each voxel of a phantom was stored as an unsigned 16-bit integer. For display purposes, we windowed the volumetric images between 5066 and 16907 and then applied a linear rescaling

to conform with the backend display functions in Psychopy, which requires 8-bit images. We utilized 160 unique 3D DBT phantoms for the search tasks described below.

Signals

Participants searched for two types of simulated lesions. The first signal was a solid sphere (0.3 mm diameter, 0.06 dva in the xy plane) akin to a small microcalcification lesion and spanned ~6 cross-sectional slices. The second signal resembled a mass lesion and was generated with a combination of several 3D ellipsoids with an average diameter of 7 mm (0.5 dva in the x, y plane). The density of the mass lesion decreased towards the edges of the signal profile, causing it to blend in with the anatomical background to a greater extent than the microcalcification signal. The mass signal spanned ~15 cross-sectional slices. Both signals were added to the background before the windowing and rescaling operations described above.

Search task

Experimental design

Human observers performed a Yes/No localization task (Abbey et al., 2018; Abbey & Eckstein, 2014) and reported whether a single signal was present or absent in the image stimulus. The experiment had three within-subjects factors, each with two levels: imaging modality (2D and 3D), CNN-CADe (searching with and without CADe support), and signal type (microcalcification and mass), totaling eight conditions. The presentation order of the levels of the first two factors was counterbalanced across participants. For example, half of the participants started the experiment with the CNN-CADe support, followed by a washout period (minimum two weeks) before they saw the same stimuli without the CNN-CADe. Of

those participants who completed the CNN-CADe conditions first, half searched through the 3D phantoms with the CADe before searching through the 2D slices of the phantoms with the CADe. The other half of the participants performed the 2D search before performing the 3D search. The same counterbalancing procedure between 2D and 3D searches was implemented for the other half of the participants who completed the search without the CNN-CADe before the washout period. The last factor, signal type, was combined into a single block of 160 trials (50% prevalence). In other words, one block contained 40 microcalcification-present trials, 40 microcalcification-absent trials, 40 mass-present trials, and 40 mass-absent trials. The presentation order was randomized across both signal type and ground truth status. Each block comprised 16 10-trial sessions with an enforced 2-minute break in between sessions to mitigate fatigue effects. Participants completed four blocks of trials, with the 2-week washout period occurring between blocks 2 and 3.

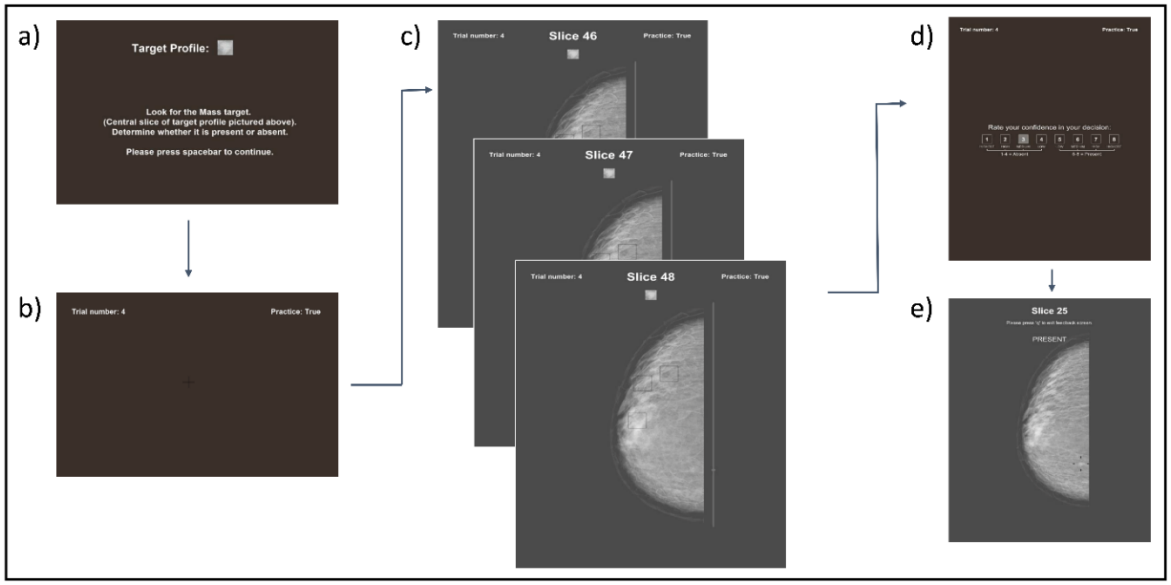

Figure 4.1. An example practice trial where a participant is looking for the mass signal in 3D, and they have access to the CNN-CADe output. a) At the beginning of each trial, the participant is informed of which signal they need to look for. In this case, it is the mass signal. A cropped image of the central slice of the mass signal is shown to the participant in addition to instructional text. b) Next, a black fixation cross is placed at a random (x, y) coordinate on top of a uniform-colored background. The participant needs to stare at the fixation cross for 1 second before proceeding to the next component of the trial. c) 3 of 64 slices of a DBT phantom are shown. When scrolling through all 64 slices, only one slice is presented on the monitor at

a time. Three cues (square bounding boxes) are superimposed on top of the image stimulus to indicate where the CNN-CADe thinks the signal is located to the participant. Note that the cues persist across multiple adjacent slices at the same (x, y) coordinate. d) After viewing the image stimulus, the participant must input a rating (1-8) indicating their confidence that the signal was absent/present in the stimulus. e) During practice trials only, the participant is shown feedback at the very end of the trial. For the 3D search task, the DBT phantom slice containing the central slice of the signal is shown on signal-present trials. A black fiducial marker surrounds the signal's location to inform the participant where the signal is located in the (x, y) plane. The central slice number is also shown at the top of the display to inform the participant where the signal was centrally located in the third dimension.

The general structure of a trial for each of the four blocks is depicted in Figure 4.1. At the beginning of each trial, participants viewed a cropped 2D image of the signal they would need to search for (Figure 4.1.a). Instructional text was also provided. The cropped image was taken from the central slice of the 3D signal profile, the cross-sectional slice corresponding to the signal's centroid. Next, they had to force-fixate a black cross placed at a random location on top of a gray background for 1 second to ensure the eye tracker was well-calibrated for the trial (Figure 4.1.b). Afterward, the image stimulus was presented to the participants until they chose to terminate the search (Figure 4.1.c). The interface of this trial component changed depending on which of the four blocks the participants were in. In the following sections, we provide a more complete description of what Figure 4.1.c entailed for each of the four blocks. Upon completion of the search component of the trial, participants rated their confidence in their decision on a scale of 1-8 (Figure 4.1.d). A rating of 1 corresponded to the highest confidence that the signal was absent, and a rating of 4 represented the lowest confidence that the signal was absent. Conversely, a rating of 5 corresponded to the lowest confidence in the signal's presence, and a rating of 8 corresponded to the highest confidence in the signal's presence.

3D search without CNN-CADe (Block I)

Upon completion of the forced fixation component of a trial (Figure 4.1.b), participants were shown the 3D image data. The "top" slice of the 3D DBT volume would always first appear on the monitor screen. Only one slice was shown to the participant at a time. The cropped image of the signal and a slice index tracker appeared at the top of the screen, above the DBT slice. This can be visualized in Figure 4.1.c. The signal template reminded participants which signal they needed to look for, and the slice index tracker displayed the current slice they were on. To view each of the 64 slices that comprised the DBT phantom image data, participants could either manipulate the mouse scroll wheel or hold and drag a widget placed on a custom-designed scroll bar located on the right-hand side of the image stimulus (also shown in Figure 4.1.c). The mouse scroll wheel allowed participants to scroll back and forth through adjacent slices. In contrast, the scroll bar provided the additional functionality of clicking on the bar to jump across multiple slices at a time.

Participants had unlimited time to perform the search and were instructed to click on the (x, y, z) coordinate that produced the most evidence of the signal's presence. They had to navigate to the slice with the highest signal contrast and click on the center of the signal profile. Clicking on the screen produced a circle (radius of 1 dva) for visual confirmation. They were instructed not to click on the screen if they did not see the signal. To end the trial, they pressed the space bar on the keyboard.

3D search with CNN-CADe (Block II)

A 3D search trial with the CNN-CADe support available mirrored the 3D search task described above but with one crucial caveat. While scrolling through the cross-sectional slices of the 3D DBT data, a varying number of square bounding boxes (3.1 dva in width/length) would appear on the screen, overlaid on top of the DBT slices. An example of what the cue

boxes looked like while the participants scrolled through 3 DBT slices is shown in Figure 4.1.c. Participants were informed that the cued locations correspond to where a computer vision model predicts the signal may be located. The cue boxes on microcalcification-present and microcalcification-absent trials persisted across 5 slices (2 slices above and below the central slice on which the box was placed), and the cue boxes on mass trials persisted across 11 slices (5 above and below the center location). The choice for the cue box locations and the number of cue boxes that appeared on screen for a given trial is described in CNN-CADe subsections below.

2D search without CNN-CADe (Block III)

Participants interacted with a single DBT slice while performing the 2D search task. Concerning Figure 4.1.c, only one DBT slice would appear on the screen, and scrolling would be disabled. Furthermore, neither the slice index tracker nor the custom scroll bar was on the monitor screen. All other aspects of the search interface depicted in Figure 4.1.c were held constant.

On signal-present trials, the DBT slice corresponding to the central slice of the signal was displayed. The image stimulus depicted in Figure 4.1.e, sans the additional feedback text and markings, provides an example of what the participants saw when searching for the mass signal in 2D. The 32nd slice of a signal-absent 3D DBT phantom was displayed to participants on signal-absent trials.

We used single DBT slices rather than simulated mammograms to model the search with a “2D imaging modality”. The noise power spectrum of 2D mammograms differs from that of single-slice DBT images (L. Chen et al., 2012). Therefore, an AI-based CADe system may have a differential impact on 2D mammograms versus single slices of DBT volumes because

image acquisition parameters and postprocessing differ across these two modalities, which can cause differences in lesion conspicuity (Horvat et al., 2019). DBT slices allowed us to isolate differences in performance across 2D and 3D searches while controlling for confounds that the image generation process may introduce.

2D search with CNN-CADe (Block IV)

The 2D search task with the CNN-CADe support replicated the 2D search task described above. However, it included cue boxes superimposed on top of the DBT phantom slice. Like in Block II, participants were informed that the cued locations represent the predicted locations made by the computer vision model for the mass/microcalcification signal.

Training and practice trials

Before completing the experiment blocks, participants partook in 4 practice blocks to familiarize themselves with the search tasks. For the 3D and 2D practice blocks without the CNN-CADe, there were 80 practice trials per block and 40 trials per signal (50% prevalence). Additionally, there were 20 practice trials (10 per signal and 50% prevalence) for the 2D and 3D search practice blocks with the CNN-CADe. These blocks were included so observers could estimate the CNN's performance and develop an internal model of incorporating its information into their decision-making process.

Block of practice trials were interleaved with each of the four experimental blocks. For example, those participants randomly assigned to the 2D CNN-CADe block completed the practice block without the CNN-CADe. This was done to help familiarize them with the overall task procedure and to develop an understanding of what the two signals looked like in a single DBT phantom slice. Then, they would complete the practice block with the CNN-CADe before starting the experimental block. Upon completion of the 2D CNN-CADe

experimental block, those same participants would complete the 3D practice block without the CADe, followed by the 3D practice block with the CADe, before starting the 3D experimental block with the CNN-CADe. Without further training, these participants would continue to the last 2 experiment blocks after the washout period.

Practice trials maintained the same design as experiment trials. However, at the end of every practice trial, feedback was given to the participants. On signal-present 2D trials, a fiducial marker was superimposed on top of the DBT slice, centered on the signal's location. If participants made a localization click on the trial, the circle centered on where they clicked was also present on the screen so they could discern where they clicked relative to the signal's location. On signal-present 3D trials, the same DBT slices were shown as in 2D (i.e., the central slice of the signal in the DBT volume) but included the slice number on which the center of the 3D signal was placed. Figure 3.1.e provides a graphical depiction of the feedback on a mass-present 3D trail. For signal-absent trials in 2D and 3D, a gray background with the text "ABSENT" was displayed to participants. Participants had unlimited time to review the feedback before proceeding to the next practice trial.

Microcalcification and mass peripheral detectability task

Upon completing the search tasks, participants partook in a forced fixation yes/no location-known-exactly detection task. We included this task to measure each participant's peripheral detectability of the microcalcification and mass signals (M. P. Eckstein et al., 2017; Lago et al., 2017; Lago, Sechopoulos, et al., 2020). Participants viewed 800 stimuli, 400 per signal. Each stimulus was a single slice of a DBT phantom sampled from a set of stimuli not shown to participants in the search task. Half the stimuli contained a signal (50%

microcalcification and 50% mass), and the other half contained no signal. The detection tasks were segregated into two separate 400-trial blocks, one block per signal.

At the beginning of each trial in a block, participants stared at a black fixation cross, superimposed on a gray background, at the center of the computer monitor. A fiducial marker was also present on the screen. The marker was centered 5 dva from the center of the fixation cross, and it appeared at 1 of 4 polar angles on any given trial: 0, 90, 180, or 270 degrees. Participants were informed that the signal would appear in half of the trials at the cued location. Therefore, they needed to only covertly attend to the marker's location and ignore all other locations in the image stimulus. After staring at the fixation cross for 1 second, the image stimulus appeared on the screen for 200 ms. The trial would abort if participants attempted to make a saccade towards the cued location. Afterward, participants encountered the same rating scale as in the search tasks (Figure 4.1.d). They had to indicate their confidence that the microcalcification/mass was present (or absent) at the cued location. In sum, we measured the peripheral detectability at 4 polar coordinates in the visual field (100 trials per coordinate and 50% prevalence).

Each block for measuring the extra-foveal processing of the microcalcification/mass signal was preceded by a block of practice trials. There were 16 trials, 2 per polar coordinate (1 signal-present trial and 1 signal-absent trial). Practice trials provided feedback at the end of each trial, similar to Figure 4.1.e.

CNN-CADe

Model overview

Our study employed an encoder-decoder U-Net CNN architecture for image segmentation (Çiçek et al., 2016; Ronneberger et al., 2015). The encoder-decoder semantic segmentation architecture allowed us to preserve a one-to-one mapping between the input stimulus size and the model output size. For our experiment, the model output, being the probability of malignancy at each voxel/pixel location in the image, is a requisite for displaying cue boxes to human observers during the search task. Furthermore, we utilized nnU-net (Isensee et al., 2019), an out-of-the-box segmentation tool built upon the basic U-Net architecture. nnU-net automates the preprocessing, network architecture, training, and post-processing configuration settings given domain-relevant information for the use case at hand (i.e., properties of the dataset, voxel spacing size, image modality, image size, etc. (Singh et al., 2020)). Moreover, nnU-net has outperformed specialized networks on various biomedical tumor segmentation tasks, demonstrating its generalizability to new datasets (Isensee et al., 2019). For this experiment, we trained and tested 4 models: microcalcification-2D, microcalcification-3D, mass-2D, and mass-3D.

Preprocessing

The input to the 3D models for training were the phantoms cropped to size 380x380x64 to improve training efficiency the non-cascade full-resolution network. The input to the 2D models for training were single slices of the phantoms of size 793x2048x1. We cropped the left-hand side of the image slices as it was a black background that provided no relevant information for training. The phantom resided on the right-hand side of the slices for all stimuli.

Network architecture

The backbone of the U-Net architecture for each of the 4 separate models consisted of an encoder and decoder module. For the encoding stage, strided convolution was used to down-sample the input spatial dimensions while increasing the feature dimensionality. For the decoding stage, up-sampling was performed using transposed convolutions, thus gradually decreasing the feature dimensionality while increasing the spatial dimensions until the output matched the input dimension size. Both the encoder and decoder were comprised of two computational blocks. Within each block, convolution operations were followed by instance normalization and a leaky-ReLU nonlinearity operation. The nnU-Net utilized a Stochastic Gradient Decent optimization to minimize the cross-entropy and maximize the dice coefficient with a preconfigured learning rate and Nestrov momentum hyperparameters set to 0.01 and 0.09, respectively. A 'polyLR' (polynomial function) regime caused the learning rate to decay across training for each parameter group.

Model	AUC	Proportion of trials cue on signal location	Connected components parameters			Number of cues			
						Signal-present trials		Signal-absent trials	
			P(malignancy) threshold	Euclidean distance (x, y) pixels	Manhattan distance (z) slices	Mean	SD	Mean	SD
3D calc	1.0	0.775	0.9	350	7	5	1.961	3.9	1.736
2D calc	0.931	0.775	0.2	350	N/A	1.1	0.304	0.825	0.549
3D mass	0.743	0.775	0.8	160	11	7.875	2.221	8.675	3.133
2D mass	0.696	0.775	0.1	140	N/A	1.150	0.662	1.375	0.807

Table 4.1. CNN-CAD model performance metrics and relevant parameters for computing the number of cues and their locations.

Training

We utilized 5-fold cross-validation for training (4 training sets and 1 validation set). Thus, a given model (e.g., microcalcification-2D) was an ensemble of 5 separate CNNs, which were later combined to make predictions in the test set. Each of the 5 constituent models was trained for 1000 epochs. One epoch for the microcalcification-3D or mass-3D model took 800 seconds to complete compared to 240 seconds for the microcalcification-2D or mass-2D models. The training was completed across 4 12 GB Nvidia GPUs. 500 cropped phantoms containing the microcalcification and 500 cropped phantoms containing the mass signal were utilized to train the 3D models. 1,500 single slices—3 slices per each of the 500 phantoms used in the 3D training set—were chosen for training the 2D models. The 3 slices per phantom corresponded to the central slice and the slices above and below the central slice.

Post-processing

After training, a given ensemble model was fed the full-resolution 3D phantom or 2D slice shown to the trained human observers. The probability of malignancy score at each voxel (3D) or pixel (2D) location was binarized using a model-specific threshold. The P(malignancy) thresholds are shown in the third column of Table 4.1 for all 4 models. Voxels above the threshold were treated as signal (1), and voxels less than the threshold were treated as background (0). We then applied a connected components algorithm—26 connectivity for 3D binarized model outputs and 6 connectivity for 2D binarized outputs—to join contiguous/neighboring signal voxels into blobs (Silversmith, 2023). This procedure resulted in multiple connected components of varying sizes per stimulus.

Testing

With the connected component output in hand, we chose the count of voxels/pixels comprising the largest component as the decision variable of the model for each of the 80 test stimuli per imaging modality and signal type combination. We assumed that the largest component would correspond to the actual signal location if present in the volume/image. Moreover, on average, phantoms containing a signal would have larger connected components than phantoms without a signal. Based on this decision variable, we computed the area under the receiver operating curve (AUC) for each of the 4 models to confirm their ability to discriminate signal from noise. The AUCs for each model can be found in the second column of Table 4.1.

Converting CNN output to CADe support tool

There are many candidate options for displaying the CNN output as a support tool to human observers. Previous studies have presented both a stimulus-level score and location-specific scores in the form of cue boxes (Pinto et al., 2021; Seah et al., 2021), saliency maps (Geras et al., 2019), or "click-to-see" model probability scores for a specific location on the image (i.e., interactive decision support (Rodríguez-Ruiz et al., 2019; Samulski et al., 2010)). We converted the connected component output into cue boxes/prompts superimposed on the image stimulus. Our central hypothesis focused on microcalcification search errors in 3D. Interactive decision support would not mitigate search errors because if the participants did not foveate the microcalcification signal, they would not click on the stimulus to activate the decision support. Interactive decision support is most helpful in mitigating decision and recognition errors, misses that occur when the observer foveates the signal but reports it as absent.

We omitted from displaying the probability of malignancy scores associated with each box because the per-pixel probability thresholds used to generate the connected components varied across the 4 CNN models. For instance, the probability of malignancy associated with cues for the microcalcification-3D stimuli would range between 0.9 and 1, whereas the probability scores would range between 0.2 and 1 for the microcalcification-2D stimuli. The difference in the range of probability scores across models would introduce information to the observer, potentially confounding our analysis.

To convert the connected components in a 3D phantom stimulus into bounding boxes overlaid on the stimulus, we first computed each component's center-of-mass coordinate (x, y, z). We then computed the Euclidean distance in the (x, y) plane between every pair of center-of-mass coordinates. We also computed the Manhattan distance between the z-coordinates for every pair of center-of-mass coordinates. We grouped components if their Euclidean distance was less than a model-specific distance threshold and their Manhattan distance was less than a model-specific threshold. Each group of connected components was converted into a single cue. The cue was placed at the mean location of all the center-of-mass coordinates in the group. The same procedure was done for the 2D connected components without the Manhattan distance calculation. The threshold parameters for this process can be found in Table 4.1, columns 5 and 6.

This grouping procedure was done to prevent overlap amongst the cued locations, which would induce visual clutter and distract from the primary task. Second, we wanted to reduce the average number of visual prompts per stimulus to less than 10 to maintain consistency with previous studies that report the number of CNN-CADe false positive prompts per image (see Table 3 in (Fan et al., 2019)). Third, we attempted to normalize the number of cues in the

signal-present and signal-absent sets of stimuli to prevent observers from utilizing this information to make their decisions. For example, in the edge case where all signal-present stimuli have at least one cue and all signal-absent stimuli have zero cues, a human observer could use the number of cues on the stimulus to determine the presence/absence of the signal.

Lastly, we equated the localization accuracy of the 4 models to ensure, from the observer's vantage point, that the CNN-CADe provided consistent, accurate information across the 2D/3D searches for the mass and microcalcification signals. The localization accuracy in 3D was defined as the proportion of signal-present trials where at least one cue prompt was less than 1.5 dva away from the centroid of the signal profile in the (x, y) plane. Moreover, the central slice of the signal needed to appear in at least one of the slices where the cue would appear on the screen. Recall that the cue boxes spanned 5 slices in z for the microcalcification signal and 11 slices in z for the mass signal. For 2D, only the former condition described for 3D needed to be met to define CNN-CADe localization accuracy.

To equate the localization accuracy across the 4 models, we applied a grid search over 3 parameters in 3D: the P(malignancy) threshold, the Euclidean distance threshold, and the Manhattan distance threshold. For the 2D models, we applied a grid search over only the first two parameters. This grid search produced a localization accuracy of 0.7775 across all models (Table 4.1, column 3). In other words, on 31 of the 40 signal-present trials, at least one cue would be placed directly over the signal.

Human performance measures and statistical analysis

We assessed overall human performance in the search tasks based on the following primary endpoints: AUC, hit rate, and false alarm rate. We supplemented this analysis by

stratifying misses into two categories: search errors and recognition errors. Furthermore, we quantified the proportion of the search area observers explored with eye movements (proportion of area covered by the Useful Field of View, or PAC UFOV) and the time participants spent searching. Lastly, for the peripheral detection task, we calculated the AUC for each signal to determine how much the mass signal was more detectable in the visual periphery than the microcalcification signal. We also combined the peripheral detectability measurements into participant-specific UFOVs (PUFOVs) to highlight individual differences for the CNN-CADe benefit in the search tasks. Below, we provide a more complete description of each analysis.

AUC-search

We employed a multi-reader multi-case (MRMC) analysis (Gallas & Brown, 2008; Obuchowski & Bullen, 2022; Roe & Metz, 1997) to evaluate significant differences in the AUC with versus without CNN-CADe decision support using the open-source MRMCaov software package available in the R programming language (B. J. Smith & Hillis, 2020). This software treats "readers" and "cases" as random effects under a generalized linear mixed effects model framework. The software also provides individual AUC estimates for each participant in each condition (with and without the CAD), assuming a binormal model. We applied this analysis 4 times, once for each search task: microcalcification 2D search, microcalcification 3D search, mass 2D search, and mass 3D search. We applied the Benjamini-Hochberg false discovery rate (FDR) correction to an $\alpha = 0.05$ level for all 4 two-tailed p-values (Benjamini & Hochberg, 1995).

In line with our primary hypotheses outlined in the introduction, we assessed whether the benefit of the CNN-CADe was significantly greater in 3D than in 2D for the microcalcification

signal. For the mass signal, we determined whether this benefit was significantly greater in 2D than in 3D. Here, we utilized a nonparametric bootstrap resampling procedure (i.e., sampling readers and cases with replacement 20,000 times) and computed the mean empirical AUC for all 8 levels of the 3 within-subject factors. For a given bootstrap iteration and signal type, we subtracted the mean AUC without the CNN-CADe from the mean AUC with the CNN-CADe for both the 2D and 3D searches. For the microcalcification signal, we subtracted the difference in AUC in 2D from the difference in AUC in 3D. For the mass signal, we subtracted the difference in AUC in 3D from the difference in AUC in 2D. We computed the proportion of difference of differences in mean AUC that were greater than 0 across all 20,000 bootstrap iterations to obtain 1-tailed p-values. In total, two p-values, one for each signal, were compared to $\alpha = 0.05$.

Hit rate and false alarm rate

Hits and false alarms were defined as ratings greater than or equal to 5 on signal-present and signal-absent trials, respectively. The number of hits divided by the number of signal-present trials (40) produced a participant-specific hit rate. The same procedure was applied to false alarms on signal-absent trials. We utilized the bootstrapping procedure discussed above (i.e., sampling readers and cases with replacement 20,000 times) to obtain differences in mean hit rate or false alarm rate for the searches with the CNN-CADe and without it. The count of bootstrapped differences in the hit rate (or false alarm rate) more extreme than 0 was divided by 20,000 and then multiplied by 2 to obtain a two-tailed p-value. We FDR corrected for 4 p-values (2D microcalcification, 2D mass, etc.) per endpoint. This nonparametric procedure, including the number of pairwise comparisons and the FDR correction, was applied to search and recognition errors, the area covered by the UFOV, and the amount of time spent searching.

Search and recognition errors

Search and recognition errors allowed us to ascertain the impact of foveal vision on detection performance by stratifying misses into two distinct categories. Search errors were defined as the subset of false negative responses where an observer failed to fixate directly on the signal. Recognition errors were defined as the complement set of misses where observers missed the signal but stared directly at it (Drew, Vo, Olwal, et al., 2013; Krupinski, 1996; Kundel et al., 1978). In the 2D search conditions, we computed the Euclidean distance between every recorded fixation position and the center (x, y) coordinate of the signal's location for a given participant and trial. If at least one fixation was at a distance less than or equal to 2.5 dva away from the signal, then the observer fixated the signal on that trial. For the 3D search conditions, we augmented the definition of fixating the signal because its profile spanned multiple consecutive slices. The Manhattan distance between the z coordinate of every fixation and the signal's central slice z coordinate was computed. If the Manhattan distance for a fixation was less than or equal to N, where N=3 for the microcalcification and N=10 for the mass, and the Euclidean distance in (x, y) was less than or equal to 2.5 dva, then that fixation was considered to be on the signal in 3D. To obtain an error rate per participant, we divided the count of each type of error by the total number of signal-present trials (40).

PAC UFOV

The proportion of the search area covered by the UFOV provides an approximate estimate of how much observers explored the 2D slices or 3D volumes with eye movements. For a given observer and trial, we "painted" a circle on all recorded (x, y) fixation locations in 2D and all recorded (x, y, z) fixation locations in 3D. The circle had a radius of 2.5 dva, the standard in the literature (Drew, Vo, Olwal, et al., 2013; Krupinski, 1996). (We include a

supplementary analysis that utilizes a signal-specific UFOV radius based on each observer's peripheral detectability of a signal). We computed the cardinality of the union set of pixels that were "painted" by the UFOV and divided this count by the number of pixels that comprised the DBT phantom slice (2D) or DBT phantom volume (3D) to obtain a proportion. We computed point estimates per observer by averaging the PAC UFOV across all signal-present and signal-absent trials.

Search time

The search time was defined as the elapsed time (in seconds) between when the image stimulus was first displayed on the monitor and when the participant pressed the spacebar to end the search component of the trial. Point estimates were obtained by averaging across all signal-present and all signal-absent trials.

AUC-peripheral detectability

We implemented a single-reader multi-case analysis to obtain participant-specific AUCs for the microcalcification and mass signals based on their rating data from the peripheral detection task. This was done using the MRMCaov software package available in R. To test whether the average AUC, across participants, was significantly lower for the microcalcification signal than the mass signal, we utilized the same nonparametric bootstrapping procedure (20,000 bootstraps) with empirical AUCs to obtain differences in mean AUCs across the two signals. A single p-value was compared to $\alpha = 0.05$. One participant chose not to complete the peripheral detection task for the mass signal, and we omitted them from this analysis.

Lastly, we used the parametric AUC estimates from the MRMCaov package to obtain participant-specific UFOV radii for each signal type (PUFOV). We assumed that detection

performance for both signals at the fovea in a location-known-exactly and signal-known-statistically task would produce AUC estimates of 1, which is a mild assumption. We then fit a half-Gaussian function:

$$AUC = \frac{\gamma}{\sqrt{2\pi\sigma^2}} e^{\frac{-E^2}{2\sigma^2}} \qquad (Eq.\ 4.1)$$

where γ and σ^2 are fitting parameters and E refers to eccentricity ($0 \leq E \leq 10$ degrees visual angle). The function was fit to two points for a single signal: ($E = 0,\ AUC = 1$) and ($E = 5,\ AUC = \widehat{AUC}$) where $\widehat{AUC}$ refers to the participant-specific estimated AUC from the MRMC model described above. We set $AUC = 0.82$ and solved for E to obtain E^*. The radius of the PUFOV for a given signal was set to E^*. An example of applying this procedure to a single subject is shown on the left-hand side of Figure 4a/b for the microcalcification and mass signals, respectively. We acknowledge that we are fitting a function with two parameters to 2 data points; thus, our fit has no error. This limitation can be fixed in future work by computing the peripheral detectability of each signal at various eccentricities.

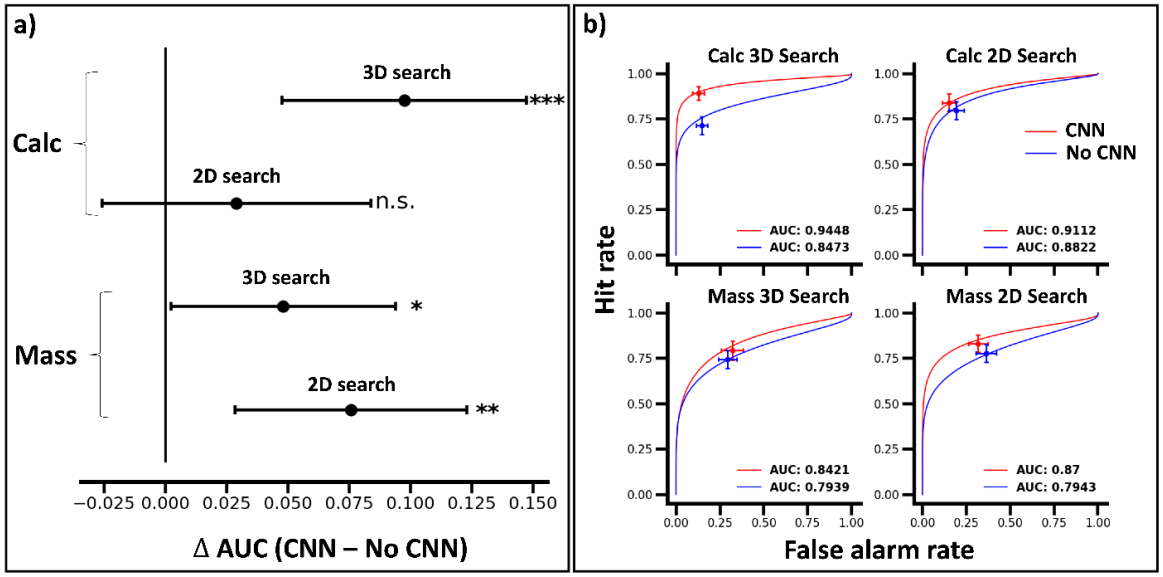

Figure 4.2. The results of the MRMC analysis depict the benefit of the CNN-CADe during the 2D and 3D searches for the microcalcification and mass signals. a) The difference in reader-averaged AUC between the

CNN and no CNN searches (scatter points) and their respective 95% confidence intervals (horizontal lines) are plotted with respect to the null hypothesis of no change in AUC when searching with versus without the CNN-CADe (vertical line centered at 0). From top to bottom, the change in AUC is plotted for microcalcification 3D search, microcalcification 2D search, mass 3D search, and mass 2D search. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and n.s. signifies $p > 0.05$. b) reader-averaged Binormal ROC curves with (red lines) and without the CNN-CADe (blue lines) are shown for microcalcification 3D search (top left), microcalcification 2D search (top right), mass 3D search (bottom left) and mass 2D search (bottom right). The area under the ROC curve is also reported for the CNN and No CNN searches in each of the 4 subplots. The scatter point in a given subplot represents the participant-averaged operating point (false alarm rate, hit rate) at the cut point 4.5, the middle of the rating scale used in our experiment. Horizontal and vertical error bars for a given operating point denote 68% bootstrapped confidence intervals (~1 standard error of the mean) for the false alarm rate and hit rate, respectively.

We recomputed the mean proportion of area covered on 2D and 3D signal-absent trials without the CNN-CADe available, using each participant's PUFOV. The area covered in signal-absent trials without the CNN-CADe provides a measure of eye movement exploration that is not confounded by finding the signal during the search or relying on the CNN-CADe prompts. The personalized UFOV, instead of the standard UFOV, normalizes peripheral detectability for a given signal across all participants by combining the two constructs into one. Next, we correlated these estimates of the PAC PUFOV with each observer's change in AUC when searching with the CNN versus without it. This correlation allowed us to ascertain whether those who explored less, while normalizing by their peripheral detectability of each signal, benefited the most from the CNN-CADe during the searches. That is, they have the most considerable change in AUC.

4.4. Results

The CNN-CADe provides the largest benefit for the 3D search of small signals

Our main objective in this study was to determine if the CNN-CADe improves performance in the 2D and 3D searches and if the benefits are contingent on what type of

signal observers had to find. Figure 4.2.a demonstrates that the change in overall search performance with the CNN-CADe in 2D and 3D depended on the signal type. For the small microcalcification signal, having the CNN-CADe available during the 3D search markedly improved the overall AUC ($\Delta\, AUC = 0.098$, 95% CI $[0.048, 0.147]$, $p = 0.0002$). However, in the 2D search, the observed change in AUC did not reach statistical significance ($\Delta\, AUC = 0.029$, 95% CI $[-0.026, 0.084]$, $p = 0.296$). Moreover, the benefit of the CNN-CADe in 3D was significantly greater than that of the CNN-CADe in 2D ($\Delta\Delta\, AUC = 0.066$, $p = 0.035$).

For the mass signal, we observed an opposite effect of the CNN-CADe on 2D versus 3D search performance. During the 3D search for the mass, the AUC change was significant but did not survive an FDR correction ($\Delta\, AUC = 0.048$, 95% CI $[0.002, 0.094]$, $p = 0.048$). On the other hand, when observers searched for the mass signal in 2D, the CNN-CADe significantly benefited their search ($\Delta\, AUC = 0.076$, 95% CI $[0.028, 0.123]$, $p = 0.002$). However, the improvement in AUC when searching with the CNN-CADe in 2D was not significantly greater than the improvement in AUC when searching for the mass in 3D ($\Delta\Delta\, AUC = 0.033$, $p = 0.133$).

The influence of the CNN-CADe on hit and false positive rates

We evaluated criterion-specific search performance measures to understand further how the CNN-CADe influenced the observer's perceptual decision-making processes. Figure 4.2.b depicts the reader-averaged ROC curves and the mean operating points (i.e., false alarm rate and hit rate pair) at the rating threshold 4.5 when observers searched with and without the CNN-CADe for the two signals in 2D and 3D.

When observers searched for the microcalcification in 3D (Figure 4.2.b, top left), the average hit rate significantly increased from 0.714 to 0.892 when the additional information from the CADe was made available to them, $p = 0.001$. We observed a modest but not significant reduction in the mean false alarm rate as well ($FAR_{No\ CNN} = 0.146$, $FAR_{CNN} = 0.127, p = 0.6844$). Similarly, when searching for the microcalcification in 2D (Figure 4.2.b, top right), the CNN-CADe increased the mean hit rate from 0.797 to 0.840, and reduced the average false alarm rate from 0.192 to 0.150. However, neither the difference in hit rate nor the difference in false alarm rate was statistically significant from 0, $p = 0.384, p = 0.317$, respectively. Together, these results suggest that CNN-CADe facilitated the detection of the microcalcification in 3D to a greater extent than in 2D and support the finding of a significantly larger change in AUC in 3D than in 2D, as discussed above.

In considering the search for the mass signal, the CNN-CADe minimally impacted the hit and false alarm rates in both the 2D and 3D modalities. During the 3D search (Figure 4.2.b, bottom left), the mean hit rate increased with the CNN-CADe ($HR_{No\ CNN} = 0.745$, $HR_{CNN} = 0.794$) but so did the false alarm rate ($FAR_{No\ CNN} = 0.293$, $FAR_{CNN} = 0.320$). However, these differences in the hit rate and false alarm rate were not significantly different from 0 ($p = 0.135$, $p = 0.624$, respectively). When observers searched for the mass signal in 2D, their average hit rate increased from 0.778 to 0.831 with the CNN-CADe, but this difference was not significantly different from 0 ($p = 0.130$). The false alarm rates were also not significantly different from one another ($p = 0.368$) despite the CNN-CADe marginally reducing the mean false alarm rate from 0.363 to 0.317. The negligible changes in hit rate and false alarm rates support the finding that the improvement in AUC in the 2D search was not significantly higher than the improvement in AUC in the 3D search for the mass signal.

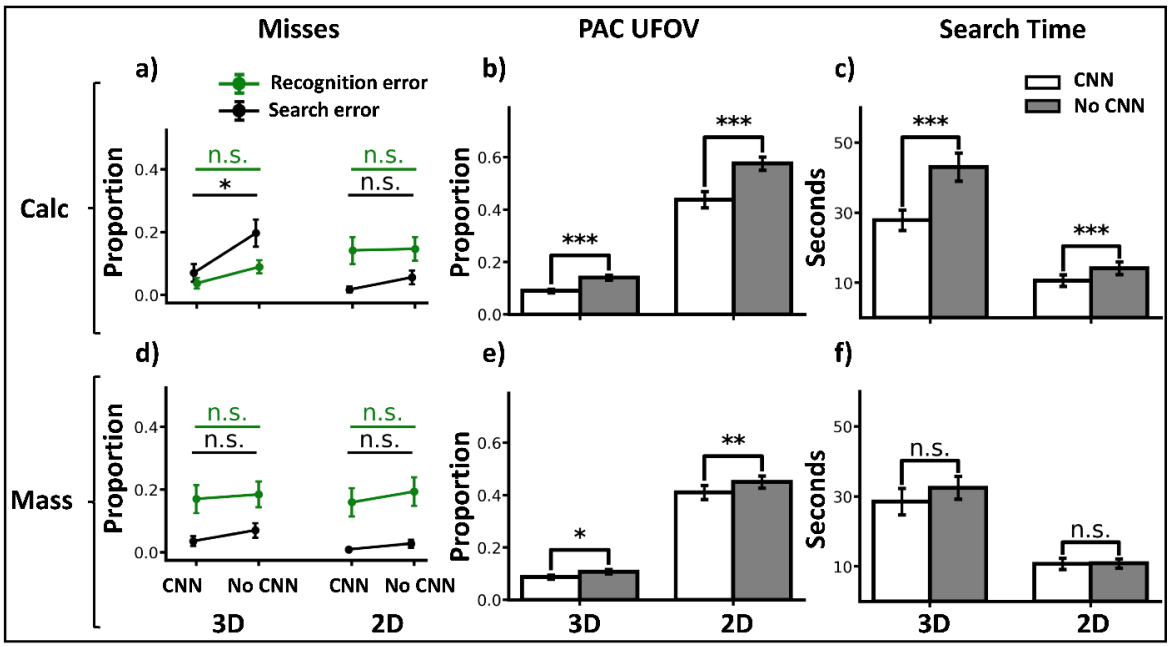

Figure 4.3. Additional measures exemplify the impact of the CNN-CADe on the participant's search strategies. a) The mean proportions of microcalcification search errors (black lines) and recognition errors (green lines) in 3D (left set of lines) and 2D (right set of lines). For a given line, the left scatter point represents a particular mean error rate when observers searched for the microcalcification with the CNN-CADe, and the right scatter point represents that same endpoint but when observers searched for the microcalcification without the CADe available. b) The mean proportion of the search area (PAC) covered by the standard UFOV (2.5 dva radius) on all trials for the 3D (left cluster of bars) and 2D (right cluster of bars) searches. White bars denote the area covered when searching for the microcalcification signal with the CNN-CADe, and gray bars represent the proportion of the area covered when searching for the microcalcification without the CNN-CADe. c) The mean search time while looking for the microcalcification signal. The breakdown of search time in 2D/3D and CNN/No CNN is kept consistent with b). d), e), and f), The same endpoints and organization of 2D/3D and CNN/No CNN data that was discussed in a), b), and c) are shown for the mass signal. All error bars represent 68% bootstrap confidence intervals (~ 1 SEM). * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$ and n.s. signifies $p > 0.05$.

The influence of the CNN-CADe on search and recognition errors

We are not only interested in whether the CNN-CADe improves search performance in 2D and 3D for both small and large signals but also in how it facilitates the detection of these signals. Our analysis of the gaze-contingent errors provides this additional insight. Figure 4.3.a depicts how search and recognition errors for the microcalcification signal in 3D and 2D changed when the bounding boxes from the CNN-CADe were made available during the searches. Figure 4.3.a left demonstrates that the mean search error rate (SER) was significantly

reduced with the CNN-CADe ($SER_{No\ CNN} = 0.197,\ SER_{CNN} = 0.070,\ p = 0.0069$). Although the mean recognition error rate (RER) was also significantly reduced from 0.089 to 0.038 when searching with the CADe, $p = 0.043$, it did not survive an FDR correction. When observers searched for the microcalcification signal in 2D (Figure 4.3.a, right), we observed that the CADe reduced the mean SER from 0.056 to 0.017 and the mean RER from 0.147 to 0.142. However, these differences were not significantly different from 0, $p = 0.060$ and $p = 0.893$, respectively.

Relative to Figure 4.3.a, Figure 4.3.d shows similar but less dramatic effects of the CNN-CADe on the search and recognition errors for the mass signal in both 3D and 2D. During the 3D search (Figure 4.3.d, left), both the mean search error rate ($SER_{No\ CNN} = 0.071$, $SER_{CNN} = 0.036,\ p = 0.125$) and mean recognition error rate ($RER_{No\ CNN} = 0.185$, $RER_{CNN} = 0.170,\ p = 0.647$) were reduced, but the changes in error rates were not significantly different from 0. When observers searched for the mass signal in 2D (Figure 4.3.d, right), the mean SER was reduced from 0.028 to 0.009, and the mean RER was reduced from 0.1938 to 0.160 when the cue boxes were present during the search. However, these differences in SER and RER were not significantly different from 0 ($p = 0.139$, $p = 0.305$, respectively).

CNN-CADe reduces eye movement exploration (PAC UFOV)

Figure 4.3.b summarizes the 2D/3D PAC UFOV when participants were tasked to look for the microcalcification signal with and without the CNN-CADe. In the the 3D search (Figure 4.3.b, left), the mean PAC without the CNN-CADe (0.140) was significantly higher than the mean PAC with the CNN-CAD (0.090), $p < 5e^{-5}$. Similarly, during the 2D search

(Figure 4.3.b, right), the PAC without the CNN-CADe available (0.576) was significantly higher than with it (0.438), $p < 5e^{-5}$.

We observed a similar trend in the PAC UFOV when observers were tasked to find the mass signal with and without the CNN-CADe in both the 2D and 3D searches (Figure 4.3.e). The PAC during 3D search (Figure 4.3.e, left) was significantly lower when searching with the CAD (0.087) than without it (0.108), $p = 0.013$. Figure 4.3.e, right, shows that the PAC in 2D was also significantly lower with the CADe (0.410) as opposed to searching without it (0.450), $p = 0.003$. In sum, regardless of the imaging modality or signal type, observers, on average, explored less of the search area with eye movements when the CNN-CADe support system was enabled.

CNN-CADe reduces the search time for the microcalcification but not the mass signal

Figure 4.3.c depicts the effect of the CNN-CADe on the time spent searching for the microcalcification in 3D and 2D. During the 3D search (Figure 4.3.c, left), there was a significant reduction in average search time from 42.994 seconds without the CADe to 27.848 seconds with it, $p < 5e^{-5}$. While performing the 2D search (Figure 4.3.c, right), observers searched for 14.141 seconds on average without the CADe and 10.606 seconds with the CADe, and this difference was statistically significant, $p < 5e^{-5}$.

Figure 4.3.f exemplifies how the CNN-CADe impacted the search time for the mass in both 2D and 3D. Figure 4.3.f, left shows a marginal reduction in search time when observers looked for the mass in 3D with the CNN-CADe (28.503 seconds) versus without it (32.445 seconds), $p = 0.226$. We also observed a slight reduction in 2D search time with the CNN-

CADe (10.751 seconds) versus without it (10.887 seconds), $p = 0.842$. However, neither the differences in 3D search time nor 2D search time were statistically significant.

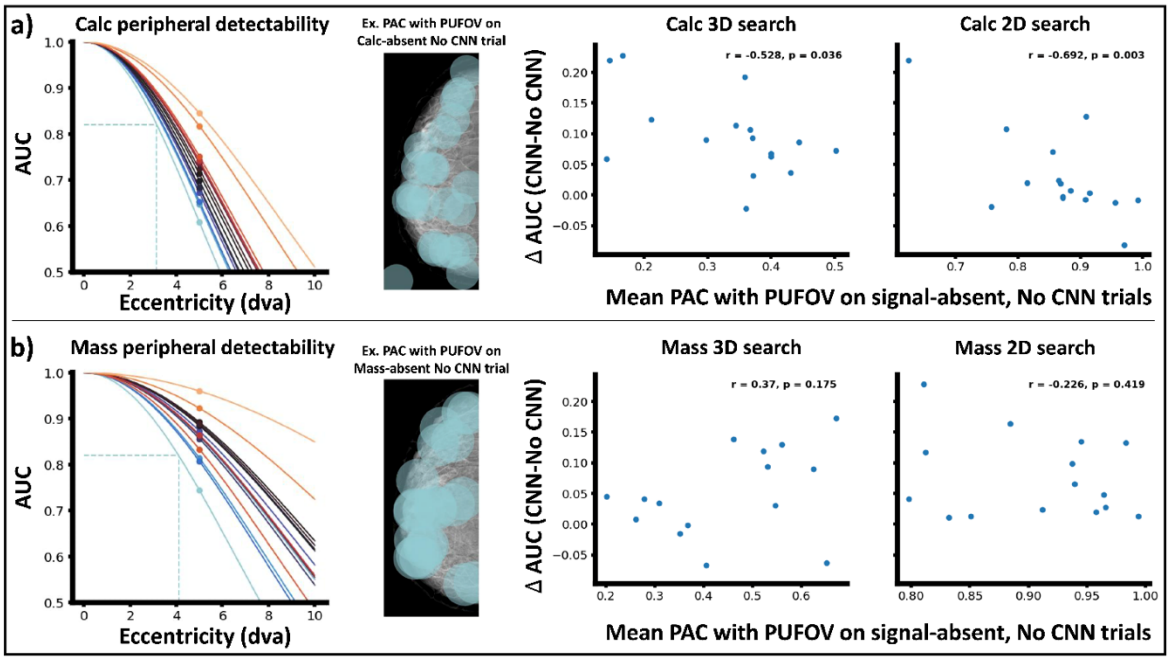

Figure 4.4. The procedure for deriving the PUFOV for each signal/participant combination. The correlation between the change in overall performance versus the mean PAC with the PUFOV on signal-absent, no CNN trials for the mass, and microcalcification signals in both 2D and 3D are also shown. a) Left, a half Gaussian is individually fit to each participant's AUC in the microcalcification peripheral detection task. Each colored line represents a fit for a different participant. Scatter points represent each participant's AUC from the forced-fixation yes/no detection task for the microcalcification signal presented at 5 dva away from the fixation point. The blue dotted horizontal line intersects the y-axis at 0.82. The blue vertical line represents the predicted eccentricity at which that participant would have an AUC of 0.82 in the peripheral detection task. This eccentricity served as the radius of the circular PUFOV for the microcalcification signal for that participant. a) Middle right, the PUFOV for the microcalcification signal derived in 4.4.a, left is "painted" on all recorded fixation positions from an example 2D microcalcification signal-absent search trial without the CNN-CADe cues available to the participant. Blue circles overlaid on the DBT slice visualize the 2D area covered by the PUFOV. a) Middle right, scatterplot relating the change in AUC (CNN AUC – No CNN AUC) during the microcalcification 3D search versus the PAC PUFOV on signal-absent microcalcification 3D trials without the CNN-CAD output available. Each point represents a single participant's change in AUC (y-axis) and how much they explored with eye movements while accounting for their peripheral detectability of the microcalcification signal (x-axis). The correlation test statistic, r, and the corresponding p-value are included in the subplot legend. a) Right, the dependent variables represented in the scatterplot in a) Middle right are visualized for the microcalcification 2D search task. b) left, middle left, middle right, and right depict the same analysis as in a) but for the mass signal.

The mass signal is more detectable in the visual periphery than the microcalcification

The scatter points in Figure 4.4.a, left, and Figure 4.4.b, left plot the peripheral detectability of the microcalcification and mass signals for all participants in this study, respectively. The mean AUC for detecting the microcalcification in the visual periphery was 0.711. In contrast, the mean AUC for detecting the mass in the visual periphery was 0.863, and this difference was statistically significant ($p < 5e^{-5}$).

Deriving signal-specific UFOVS for each observer (PUFOVS)

Figure 4.4.a, left, also includes the half-gaussian fits to each observer's peripheral detectability of the microcalcification signal. Figure 4.4.b, left, shows the same type of fits but for the mass signal. The half-gaussian fits were used to derive signal-specific UFOVs (PUFOVs) for each participant at an AUC threshold of 0.82. The middle left subplots of Figure 4.4.a and Figure 4.4.b depict the process of "painting" the PUFOVs on all recorded fixation positions from a single participant during the 2D searches of the microcalcification and mass signals, respectively. These fixations were obtained from two searches on signal-absent trials without the CNN-CADe output available. Of note for the participant's data shown here is that the radius of the circular PUFOV for the mass signal is larger than the radius of the PUFOV for the microcalcification signal. Thus, when incorporating the PUFOV into the computation of the PAC, more of the DBT phantom slice is covered in the mass trial than in the microcalcification trial.

Individual differences in the AUC benefits of the CNN-CADe correlate with PAC PUFOV

Our last analysis evaluates whether those who explore less in 2D/3D with their PUFOV benefit the most from the CNN-CADe. Figure 4.4.a, middle right exemplifies this relationship

for the microcalcification 3D search. We observed a strong negative linear correlation between how much people explore with eye movements and their change in AUC when searching with the CADe versus without it ($r = -0.528$, $p = 0.036$). In short, those who tended to explore less of the 3D volume with eye movements benefited the most from the CNN-CADe.

The correlation analysis reported above depends on two free parameters: 1) the half-gaussian function we chose to fit the peripheral detectability data to, and 2) the AUC threshold of 0.82 for computing the radius of the PUFOVs for each observer. Therefore, we reran the analysis using AUC thresholds ranging from 0.82-0.9 in steps of 0.02. We also fit a line rather than a half-gaussian to the peripheral detectability estimates and used the same range of AUC thresholds. Across the set of AUC thresholds and the two fitting functions (10 models), the correlations ranged from -0.528 to -0.416 (mean $= -0.4793$, std $= 0.0463$).

We also ran this same analysis for the 2D search of the microcalcification (Figure 4.4.a, right). Like the microcalcification 3D search, we observed a strong negative linear relationship between these two variables ($r = -0.692$, $p = 0.003$). Furthermore, considering the range of AUC thresholds and two fitting functions, the correlations spanned from -0.692 to -0.607 (mean = -0.6624, std equals $.0286$). However, caution should be taken in interpreting the strength of this negative linear relationship because one participant was an outlier (Figure 4.4.a, right, top left corner). Removing this person from the analysis produced a correlation of -0.373 ($p = 0.171$) at an AUC threshold of 0.82 for the half-gaussian fitting function.

We ran the same analysis for the mass signal in the 2D and 3D searches to understand if this relationship holds across both small and large signals. For the mass 3D search (Figure 4.4.b, middle right), we observed a positive linear relationship between the mean PAC with

the PUFOV on signal-absent trials without the CNN-CADe and the change in AUC between searching with versus without the CNN-CADe. However, this relationship was not statistically significant ($r = 0.370,\ p = 0.175$). Correlations ranged from 0.185 to 0.448 (mean $= 0.3302$, std $= 0.075$). For the mass 2D search (Figure 4.4.b, right), we observed a negative linear relationship between these two dependent variables, but this relationship was also not significant ($r = -0.226,\ p = 0.419$). Here, the correlations ranged from -0.226 to -0.114 (mean $= -0.169$, std $= 0.047$).

4.5. Discussion

Our main objectives in this experiment were to assess 1) how the benefits of the CNN-CADe vary across 2D and 3D searches and 2) how the support system interacts with the size of the searched signal across these two imaging modalities. Our results show that the CNN-CADe brings about added benefits for the 3D search of small signals, and to a lesser extent, it improves the 2D search of large signals. To better understand this nuanced interaction, we quantified how much the CNN-CADe mitigated the microcalcification and mass search and recognition errors across the 2D and 3D modalities.

For example, at the outset, we hypothesized that the CNN-CADe provides unique benefits to the 3D search of the small microcalcification signal by guiding an observer's eye movements to suspicious locations cued by the model observer, effectively reducing search errors. Recall search errors result from the interaction between under-exploring the 3D volumetric data with eye movements and having low detectability of the microcalcification signal in the visual periphery (Lago et al., 2017, 2018, 2019; Lago, Jonnalagadda, et al., 2021). Conversely, when searching for the microcalcification signal in 2D, it is relatively

easy to direct one's center of gaze to most regions of the DBT slice in a time-efficient manner. Thus, we predicted that the benefit of the CNN-CADe would be less pronounced in this case because extra-foveal processing would have a diminished influence on search errors. Indeed, our results show that the CNN-CADe markedly reduced search errors in 3D but not in 2D (Figure 4.3.a). These results are commensurate with the fact that the CNN-CADe induced a significant increase in 3D search AUC but only a marginal improvement in 2D search AUC (Figure 4.2.a). Moreover, the difference in AUC for the 3D search was significantly higher than in the 2D search.

We can ascertain that the participants relied heavily on the marked locations made by the CNN-CADe when searching for the microcalcification signal because not only did the search error rate decrease when the model output was available, but participants, on average, explored less with eye movements (Figure 4.3.b). They also searched for a shorter period when the cued locations were available (Figure 4.3.c). Thus, if an observer adopts a search strategy focusing on visually inspecting the cued locations and the cues are highly accurate, we expect observers to explore less and for a shorter duration while maintaining high sensitivity and specificity (Deza et al., 2019). These findings highlight the importance of having an accurate auxiliary aid when performing life-critical tasks such as early cancer screening. Prior work has shown that inaccurate CADe systems can increase misses because observers explore less of the search space with eye movements when it is made available to them (Drew et al., 2012), a possible consequence of *automation bias* or overreliance on the machine (Alberdi et al., 2004). This effect can be particularly pernicious as 3D imaging modalities become the standard of care for breast cancer detection.

Our second hypothesis posited that the CNN-CADe would benefit the detection of the mass signal in 2D to a greater extent than in 3D. The mass signal is more detectable in the visual periphery than the microcalcification signal (Figure 4.4.a, left versus Figure 4.4.b, left), and it spans many more slices in 3D than the microcalcification signal. Thus, more signal information in 3D can be integrated. However, in 2D, the simulated glandular and adipose tissue in the DBT slice can obfuscate or visually mask the mass signal (Mello-Thoms et al., 2003, 2005). The signal profile is dominated by low spatial frequency information, and there is high energy at low spatial frequencies in the noise power spectrum of the DBT phantom slice (L. Chen et al., 2012). In sum, we expect more recognition errors and false positives in 2D than in 3D, and the CNN-CADe should mitigate these 2D errors.

Our results partially align with this hypothesis because the CNN-CADe improved the overall search AUC for the mass in 2D but not in 3D (Figure 4.2.a). However, the improvement in 2D was not significantly greater than in 3D. Despite observing a significant improvement in AUC for the 2D search with the CNN-CADe, we did not find a significant reduction in false alarms (Figure 4.2.b, bottom right) or recognition errors (Figure 4.3.d, right). Interestingly, observers explored less of the 2D DBT slice with the CADe available (Figure 4.3.e) but did not spend significantly less time searching (Figure 4.3.f). One interpretation is that in the presence of a "second opinion," participants spent more time scrutinizing only the cued locations.

An important finding from this work is the considerable inter-observer variability in the benefits of the CNN-CADe on 3D search for small signals. Moreover, this variability can be related to an observer's exploration behaviors, quantified via the PUFOV. This claim is realized by our analysis of individual differences (Figure 4.4.a, middle right), which

demonstrates a negative correlation between the change in AUC when searching with the CADe (versus without it) and the mean proportion of the search area covered with the PUFOV on signal-absent trials with no CNN-CADe. This finding suggests that those who explored less of the 3D DBT phantom benefited the most from the cued locations. Additionally, our PUFOV construct considers the peripheral detectability of the microcalcification signal. Therefore, if an observer makes many eye movements during the search but has poor peripheral detectability, they should still benefit from the CNN-CADe because their poor peripheral vision will reduce their effective eye movement exploration.

On the other hand, we did not observe a significant correlation between the PAC with the PUFOV and the change in AUC when participants searched for the mass signal in 2D or 3D (Figure 4.4.b, middle right and right). The detection of the mass signal, with signal location uncertainty, is noise-limited (Burgess et al., 2001). That is, observers should not be influenced by extra-foveal processing and eye movement exploration but rather by signal contrast and how this attribute interacts with the anatomical noise embedded in the DBT phantom. In this regard, properly placed CADe prompts on suspicious locations scrutinized by the observer may induce increased confidence in their decision because the model reassures the observer's initial suspicion.

Our study has inherent limitations worth enumerating to help contextualize our results within the broader medical imaging field. First, we utilized trained human observers as opposed to radiologists. Given the data-intensive nature of our experiment, we opted to run non-expert observers because it allowed us to run many trials while collecting eye-tracking data. As a consequence, the external validity of our findings may be limited in scope because expertise mediates observer performance (Nodine & Mello-Thoms, 2000; Waite et

al., 2019). Despite differences in performance due to expertise, studies have shown how bottlenecks and properties of the visual system common to naïve and radiologist observers (Krupinski, 2010) result in similar effects across the two cohorts (M. Eckstein et al., 2003; Lago, Jonnalagadda, et al., 2021; Wolfe et al., 2016). For instance, (Lago, Jonnalagadda, et al., 2021) demonstrated that trained human observers and radiologists are similarly susceptible to making search errors when tasked to find microcalcification-like signals in 3D volumetric images. This finding can be explained by the neurophysiological constraints of the human visual system rather than expertise. Hence, we would expect a CNN-CADe system to aid the detection of microcalcifications in DBT volumes within a clinical setting.

Another limitation of our study concerns how our 3D search task differs from how 3D volumetric images are interpreted in a clinical setting. When radiologists interpret DBT data, they have available either a 2D mammogram or a 2D synthetic (2D-S) view generated from the DBT data. Prior work has demonstrated how a 2D-S can guide eye movements in 3D and thus mitigate search errors (D. S. Klein et al., 2023). Therefore, the presentation of CADe prompts may provide redundant information that can otherwise be extrapolated from the 2D-S. The CNN-CADe may only provide radiologists with little additional benefit beyond a reduction in reading time (Uematsu et al., 2023).

The signals used in our study pose additional caps on the external validity of our findings. First, observers knew which signal to search for at the outset of every trial. However, radiologists are not privy to this information in practice—there is signal uncertainty. Therefore, radiologists must maintain multiple signal templates in memory while examining medical images. Second, microcalcifications often appear in clusters. However, here, we had observers search for a single microcalcification. However, prior

work has shown that microcalcification clusters have low peripheral detectability (Lago, Sechopoulos, et al., 2020). Our results suggest that the benefits of the CNN-CADe in 3D search may extend to microcalcification clusters. Third, there are signals radiologists screen for indicative of malignant lesions, particularly architectural distortions, that were not investigated in this study. Despite the differences between our experimental design and what is observed in clinical practice, future studies with radiologists (or radiology residents) can measure their peripheral detectability of various signals and quantify how much they explore 3D volumetric images with eye movements. Our work provides specific predictions about how a CNN-CADe would benefit the 3D search with those measurements in hand.

Lastly, cognitive factors such as fatigue (Reiner & Krupinski, 2012) and criterion shifts that arise from low target prevalence rates in cancer screenings (Wolfe et al., 2007) might interact with the search effects in this study. Similarly, our study included one signal per case, not addressing instances with multiple lesions, often leading to the satisfaction of search (Berbaum et al., 1990; Fleck et al., 2010; Tuddenham, 1962).

4.6. Conclusion

Recent advances in artificial intelligence-based computer-aided detection algorithms can improve human observer search performance in 3D volumetric medical images where interpretation time and effort far exceed the visual examination of 2D medical images. Our study suggests that CNN-CADe brings about greater performance benefits to the 3D search of small signals (vs. 2D search) by reducing search errors caused by the under-exploration of the volumetric data. Our proposed methodology for measuring observer 3D search under exploration has the potential to identify individuals who would benefit the most from the CNN-CADe support system.

V. More than meets the (single) eye: the greater benefits of group decision-making for visual search in large 3D volumetric medical images

5.1. Abstract

The hallmark of a group's collective intelligence is its superior decision-making performance compared to the average performance of groups. This longstanding and replicable phenomenon across various perceptual tasks is reflected in the fact that certain countries require independent double reading for early cancer screening. Interpreting medical images is difficult, causing considerable variability in radiologists' performance, but it is often the case that two heads are better than one. Compounding the task's difficulty is that radiologists are beginning to visually scrutinize large 3D volumetric medical images instead of more traditional 2D displays. An interaction between eye movement under-exploration of the 3D data and the foveated nature of the human visual system can cause trained observers and radiologists to miss small lesions that are hard to detect in the visual periphery. To test the theoretical benefits of wisdom of crowds for 3D imaging modalities, twelve trained observers (Experiment 1) searched through Digital Breast Tomosynthesis (DBT) phantoms (3D search) and single slices of the DBT phantoms (2D search) for a small microcalcification signal and large mass signal. We show that a simple averaging of group member's confidence scores and an asymmetric maximum-confidence slating rule, which

changes the majority's decision if at least one member provides the highest signal-present confidence rating, boosts the group's performance to a greater extent than (1) a majority vote decision rule and (2) the mean performance of the group. Moreover, we identify that these signatures of the wisdom of crowds for visual search tasks are uniquely enhanced for detecting the microcalcification in the 3D DBT phantoms, suggesting that the group can counteract the decrement in individual search performance caused by under-exploration. In Experiment 2, twelve radiologists searched for the microcalcification and mass signals in the 3D DBT phantoms. We show that the expected benefits of group decision-making for 3D search generalize across expertise levels. In particular, the average and majority vote with exception pooling models outperform the majority vote rule. These findings provide new theoretical insight into the collective intelligence of groups for complex visual search tasks such as interpreting large 3D volumetric images.

5.2. Introduction

Many societally important actions, from jury verdicts (Tiley, 1969) to financial forecasting (H. Chen et al., 2014) to early cancer screening (Taylor-Phillips & Stinton, 2019a), rely on decisions as a group or by comparing independent judgments between individuals. Group decisions are commonly formulated by taking a majority vote across binary choices or computing a weighted average over numerical estimates. Group decision-making can often attain higher performance relative to its members—including the group's best performer—in several perceptual tasks such as estimation (Galton, 1907; Merkle & Steyvers, 2011), prediction (Hueffer et al., 2013; Kattan et al., 2016), and detection (Brennan et al., 2019; Kurvers et al., 2016; Wolf et al., 2015). This benefit of aggregating

individual judgments, colloquially referred to as the Wisdom of Crowds effect, presupposes a diverse set of individuals whose judgments are independent of one another (Surowiecki, 2005).

There has been a long history of trying to understand the upper bound (optimal) of the task-accuracy benefits that arise from aggregating judgments. Classic work has utilized Signal Detection Theory (SDT) to make upper-bound predictions for detection and classification tasks (Green, 1966; Green & Swets, 1989; Sorkin & Dai, 1994). SDT models assume each individual's judgment on a trial is based on an internal variable sampled from one of two normal distributions (binormal model), one for the signal-absent trials and another with a higher mean for the signal-present trials. Each individual's sensitivity is described by the distance between the two distribution means in standard deviation units (d'). This modeling framework has been successful at explaining how the group outperforms its members in simple perceptual tasks ranging from visual search (M. P. Eckstein et al., 2012; Juni & Eckstein, 2017; Saha Roy et al., 2021) to discriminating ruler-like stimuli (Sorkin et al., 1998, 2001) and for explaining how the majority vote rule often approximates the optimal combination rule.

But does this modeling framework correspond to all decision-making circumstances that groups might encounter? The classic SDT model assumes that individuals with high or low d' are well-described by their two associated normal distributions (and d') for every trial/decision. In practice, there are tasks in which, from decision to decision (trial to trial), the probability of making a correct response (and d') might vary—the individual might have a higher probability of making a correct decision in some trials and a lower probability in other trials depending on the circumstances (Prelec et al., 2017). Consequently, from the

group's perspective, the individual with the highest probability of a correct decision (d') might vary from trial to trial. Studies have shown that these scenarios result in greater accuracy benefits from pooling individual judgments (Juni & Eckstein, 2015) and group decisions (Juni & Eckstein, 2017). In addition, the commonly deployed and usually effective majority vote rule (Hastie & Kameda, 2005) can become highly suboptimal or even perform worse than an averaging of judgments. These effects can be captured by an extended SDT model (SDTmix) that assumes, for each individual, a sampling of an internal variable from a mixture of normal distributions on a trial-by-trail basis (Juni & Eckstein, 2017).

What real-world situations might lead to these circumstances? One example would involve a panel that is given a battery of questions spanning multiple knowledge domains and a scenario in which, from decision to decision, a varying minority of panelists often have high expertise and express very high confidence in their decision. Another example is visual search, in which observers try to find a target in a large image in a limited time. In such searches, a target may be difficult to see in the visual periphery, and the limited search time may preclude the observer from exhaustively exploring each image region with eye movements. Consequently, an observer will only fixate the target on a subset of trials depending on their particular eye movement scan path, which will, in turn, impact their trial-to-trial d'. At the group level, a different individual on each trial will have the highest probability of target detection (Juni & Eckstein, 2017).

What remains unknown is whether any real-world scenarios might also show these greater benefits of the wisdom of crowds. In medical image perception, radiologists are tasked with searching through images to screen for early signs of cancer. Moreover, radiologists are now visually scrutinizing large volumetric images produced by 3D imaging

modalities (Health, 2023; Smith-Bindman et al., 2008; Williams & Drew, 2019). Several eye-tracking studies with both radiologists and trained human observers have shown that humans fail to exhaustively scan with eye movements the large set of cross-sectional slices that constitute a 3D volumetric image (Drew, Vo, Olwal, et al., 2013; Lago, Jonnalagadda, et al., 2021; Rubin et al., 2015). In the case of breast cancer screening, small microcalcification signals are hard to detect in the visual periphery (Lago, Sechopoulos, et al., 2020) and can often go undetected during 3D search due to eye movement under-exploration (D. S. Klein et al., 2023; Lago, Abbey, et al., 2020; Lago, Jonnalagadda, et al., 2021). Taken together, 3D visual search for small microcalcification signals satisfies two preconditions enumerated by SDTmix—low peripheral target detectability and low eye movement coverage of the image—warranting further investigation as to whether or not the expected benefits of group decision-making extend to this type of real-world visual search task.

In this study, we ask the following questions. What are the quantitative performance benefits of group decision-making in 3D search? Do these benefits depend on the type of signal (small or large) observers are tasked with looking for? Are these performance benefits similar to what would be expected in traditional 2D search tasks? Moreover, how do different decision rules or pooling algorithms compare in 2D and 3D searches?

To answer these questions, in experiment 1, trained naïve observers searched with no time constraints through 3D digital breast tomosynthesis (DBT) phantoms (3D search) and single slices of the DBT phantoms (2D search) for a small microcalcification-like signal and a large mass-like signal. We also measured each observer's peripheral detectability of the two signals using a forced-fixation yes/no detection task with the signals placed at 5 degrees

of visual eccentricity from a fixation point. Search performance and eye movement exploration were quantified for all four search tasks and considered with the peripheral detection task to ascertain whether the two preconditions in SDTmix were met.

Next, the relative efficiencies of three pooling algorithms were computed for group sizes ranging from 2-9 members in all four search tasks. The relative efficiency is defined as the squared ratio in d' between a pooling algorithm (numerator) and a statistically optimal decision maker (denominator) (Tanner Jr & Birdsall, 1958). Under the statistically independent model of observers, an upper bound in performance benefits of pooling judgments across group members is determined by the sensitivity of the ideal group—a hypothetical, statistically optimal group—which can be computed by optimally integrating the individual d' estimates for each group member (Sorkin & Dai, 1994).

The pooling models we investigated in this work are the average (AVG), majority vote (MAJ), and majority vote with exception (MAJe). The last algorithm was included because it implicitly captures the notion of sampling from a mixture of Gaussians on a trial-by-trial basis described in the SDTmix framework. Relative efficiencies for each algorithm were contrasted across the four searches. We also quantified differences in relative efficiencies between search tasks for a given pooling method.

In a second experiment, radiologists performed the 3D search for the same small and large signals on a subset of the stimuli the naïve observers saw in Experiment 1. The same group decision-making analyses were performed on the radiologists to determine if our results generalize across expertise levels.

Based on this experimental framework, we hypothesize that the AVG and MAJe, but not the MAJ pooling algorithm, will incur the expected benefit predicted by SDTmix for the

small microcalcification signal in the 3D search. In particular, we predict that the relative efficiencies of these two pooling methods should be higher than the MAJ pooling method. Additionally, we hypothesize that the expected performance benefits of the AVG and MAJe pooling algorithms will be higher in the 3D search of the microcalcification signal than in the analogous 2D search for the same signal. Thus, we predict that the relative efficiencies in the 3D search will be higher than in the 2D search for these two pooling models.

The large mass-like signal is more detectable in the visual periphery than the small microcalcification signal. Therefore, we hypothesize that predictions outlined in the SDTmix framework will not apply to the 3D search of the mass signal. Specifically, we predict that the relative efficiencies across the three pooling algorithms will be similar. Similarly, for a given pooling method, we expect no differences in relative efficiency across the 2D and 3D searches for the mass signal.

5.3. Experiment 1

5.3.1. Methods

Participants

Twelve undergraduate students (58% female, age range 18-22) from the University of California, Santa Barbara, were recruited for this experiment. All twelve observers provided informed written consent (protocol # 12-23-0301) and received course credit for participation. All observers maintained normal or corrected-to-normal vision throughout the duration of the experiment.

Apparatus

Participants viewed stimuli in a darkened room (2 lux) on a medical grade grayscale 5.8 MP DICOM monitor (MDNG-6121 Barco) with a screen resolution of 2096x2800 pixels.

The refresh rate was 24 Hz, and the screen dimensions were 325x430 mm (x, y). Participants sat at a viewing distance of 750 mm, which translates to 45 pixels per degree of visual angle (dva). An eye tracker (SR Research Eyelink Desktop Mount) was positioned 600 mm from the chinrest and sampled the participant's gaze position at 2000 Hz. Before the experiment began, participants needed to pass a 9-point calibration and validation procedure with an average error during validation of less than 1 dva and a max error of less than 1.5 dva across the nine tested points on the grid. Fixation events and Saccades were collected using velocity and acceleration thresholds of 30 deg/s and 9,500 deg/s^2, respectively. The experiment was designed in PsychoPy (Peirce et al., 2019), a Python programming package utilized for psychophysics.

Stimuli

Phantoms. The current study focused on detecting cancer-like lesions embedded in 3D DBT phantoms and 2D slices of the corresponding 3D phantoms. The OpenVCT virtual breast imaging software (Bakic et al., 2018; Pokrajac et al., 2012; Predrag R. Bakic, 2017) enabled us to simulate the spatial structure and relationship of anatomical tissues like the skin, Cooper's ligaments, and adipose and glandular tissue (prevalence of 15%-25%) which are visible in actual DBT volumes. Each 700 ml simulated phantom was compressed in the mediolateral direction at 6.33 mm thickness, and the reconstruction parameters were chosen for 100 μm in-plane resolution and 1 mm depth sampling. The resultant 3D voxel arrays were of size 822x2048x64, and each voxel was stored as an unsigned 16-bit integer. We windowed each voxel between 5066 and 16907 and then linearly rescaled the values to conform to the requirements for the display software. A single 3D voxel array was stratified into 64 cross-sectional slices. Each slice was a 2D image that subtended 18.3x45.5 dva

(822x2048 pixels). In total, 160 DBT phantoms were utilized in the search tasks described below.

Signals. The OpenVCT software allowed us to insert lesions at random (x, y, z) locations within the confines of the DBT tissue but not near the edges of the stimulus. This was done before windowing and rescaling. There were two lesions: one mass-like lesion and one microcalcification-like lesion. The mass was modeled as a combination of multiple 3D ellipsoids with an average diameter of 7 mm. The density of the lesion decreased gradually away from the object's centroid. As a result, its geometric profile blended with the anatomical background surrounding it, losing contrast towards the object's edges. The signal spanned across multiple cross-sectional slices but generally was only visible within +- 10 slices in either direction from its center z-coordinate. The microcalcification signal was modeled as a solid sphere with a 0.3 mm diameter and spanned 7 slices in the z dimension. It appeared on a single cross-sectional slice as a high-contrast rod-like “pixel.”

In the following subsections, we use the term “central slice.” The central slice should not be confused with the 32nd cross-sectional slice in the DBT volume. The central slice refers to 1 of the 64 cross-sectional slices on which a signal was inserted into the 3D DBT volume. It is a planar view of the centroid of the signal in 3D and provides the most visual evidence (highest signal contrast) of the signal. For example, if the microcalcification signal were inserted into the DBT volume on the 44th slice at a particular (x, y) coordinate, it would be visible on only the slices ranging from 41 to 47, and the central slice would be 44. There is no central slice for signal-absent DBT phantoms that do not contain either signal.

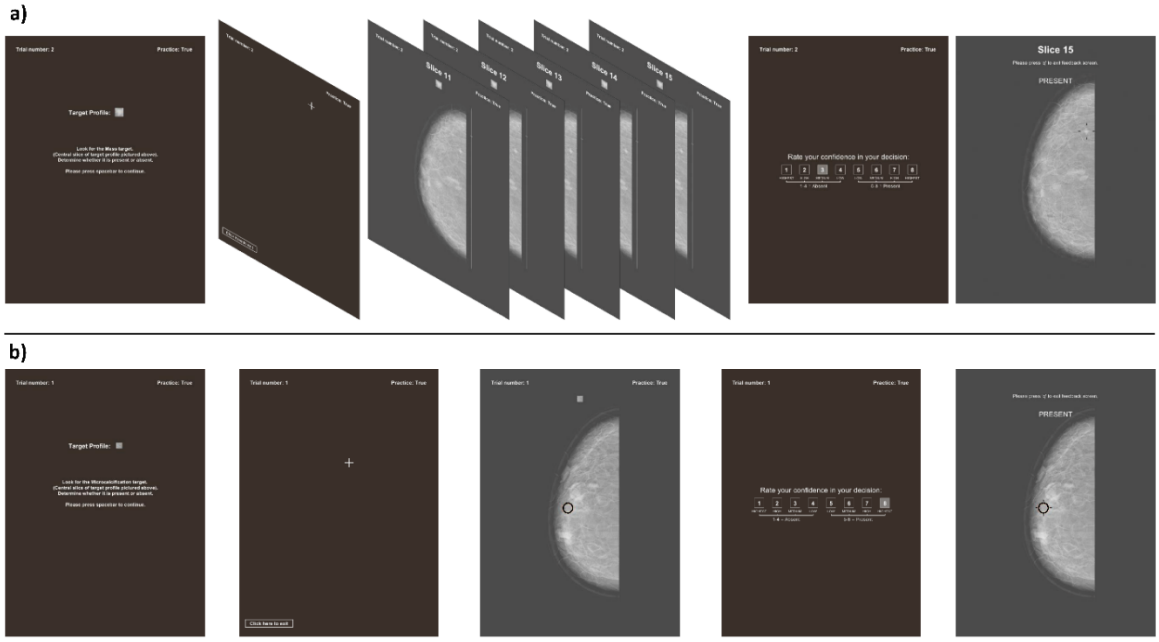

Figure 5.1. Trial flow diagram for the 3D and 2D search tasks. a) Depiction of a mass-present 3D search practice trial. From left to right, the participant was first notified which signal to look for. Next, a fixation cross would appear on the screen at a random location, and the participant needed to fixate on the cross for 1 second before proceeding to the image stimulus presentation portion of the trial. The fixation cross here is white for display purposes only. Once the image stimulus appeared on the screen, the participant could scroll through different slices. We display 5 of 64 slices, two above and two below the central slice. After completing the search portion of the trial, the participant had to indicate their confidence in the presence or absence of the signal. Lastly, they were presented with the central slice of the DBT volume that contained the signal. A fiducial marker was placed around the signal to inform the participant of its (x, y) location. b) Depiction of a microcalcification-present 2D search practice trial. From left to right, the participant encountered the same sequence of events in the 2D trial as in the 3D trial. However, only one cross-sectional slice was shown to the observer when the image stimulus appeared on the screen.

Search tasks

Overview. Participants performed a yes/no localization task. This was a 2x2 within-subjects experimental design. The first factor was imaging modality, which had two levels: 3D search and 2D search. The second factor was the signal type, with two levels: microcalcification and mass. The first factor was blocked and counterbalanced across participants—half the participants completed the 3D search task before completing the 2D search task, and the other half completed the 2D search task before completing the 3D search task. The signal type was intermixed within each block so that observers would look

for the mass signal or the microcalcification signal on any given trial, but not both simultaneously.

Each person completed 160 trials per experimental block, totaling 320 trials. Within a block, 80 trials contained one of the two signals but not both. The other 80 trials contained neither signal (i.e., 50% target prevalence). In other words, the participants saw 40 microcalcification-present, 40 microcalcification-absent, 40 mass-present, and 40 mass-absent trials. The trial presentation order was randomized across signal type and ground truth status.

Each experimental block was broken into 16 mini-blocks, with ten trials per mini-block to avoid fatigue effects. Participants were encouraged to take short breaks between mini-blocks and complete as many as possible within a 2-hour session. The 3D search task required multiple 2-hour sessions across several days, 4-5 days on average. The 2D search task also required multiple 2-hour sessions, but participants completed all trials over two days, on average. Below, we provide a complete description of the trial flow of the 3D and 2D searches.

3D search. Figure 5.1.a depicts the general task procedure for a single trial in the 3D search condition. At the beginning of each trial, a cropped 2D image (64x64 pixels) corresponding to the central slice of one of the two signals was displayed to the participant (Figure 5.1.a, left). Additional instruction text was also present on the screen. After pressing the spacebar key to acknowledge they needed to look for that signal, a black fixation cross appeared at a random location on top of a neutral gray background (Figure 5.1.a, middle left). Participants were required to stare at the fixation cross for 1 second to proceed to the search component of the trial. We included this procedure to ensure that the eye tracker was

adequately calibrated at the beginning of each trial (i.e., custom drift check). Participants were allowed to recalibrate the eye tracker during this trial phase if needed.

After successfully staring at the cross for 1 second, the image stimulus would appear on the screen. Given the 3D nature of the DBT image data (Figure 5.1.a, middle), participants did not see the entirety of the stimulus at once. Instead, they viewed a single cross-sectional slice of the 3D volume at a time on the computer monitor. In total, 64 slices required visual inspection. Participants would begin the search by inspecting the slice at the top of the image stack. They could freely scroll back and forth through all 64 slices, at their own pace, by manipulating the mouse scroll wheel. They had unlimited time to perform the search.

In addition to the image stimulus, the monitor screen displayed a slice index tracker, the 2D cropped image of the signal they needed to look for, and a custom-designed widget scroll bar. These features are shown in Figure 5.1.a, middle. The slice index tracker and 2D cropped image of the signal were displayed above the image stimulus. The slice index tracker indicated to the participant which of the 64 images was currently being displayed on the monitor. The 2D cropped image reminded participants which signal they needed to look for. The scroll bar appeared to the right of the image stimulus. The scroll bar tracked the slice number and visually indicated where the participant was currently searching in the 3rd spatial dimension. It consisted of a horizontal bar superimposed on top of a vertical bar. The vertical position of the horizontal bar moved in concert with the slice index tracker as the participant scrolled through the 3D volume. Participants could also click on the vertical line of the scroll bar to jump across slices. For instance, if they were on slice 64 and wanted to return to the top of the image stack, they could click the top of the scroll bar to return to slice 1.

Participants had two options to end the search. If they found the signal, they had to click on the image stimulus at the (x, y) location where they believed the signal was present. Moreover, they had to navigate to the slice that provided them with the most visual evidence of its presence, presumably the central slice. Once they clicked on a location, a circle would appear at the marked area for visual confirmation. They were allowed to remove clicks or click at another location but were instructed to mark at most one location per trial. Afterward, they would press the spacebar to end the search portion of the trial. If they did not find the signal on the trial, they did not click anywhere and pressed the spacebar to end the search.

Next, participants had to rate their confidence in their decision on a scale of 1-8 (Figure 5.1.a, middle right). Confidence ratings of 1-4 mapped to signal-absent decisions. A rating of 1 represented the highest confidence that the signal was absent, and a rating of 4 represented the lowest confidence that the signal was absent. Conversely, a rating of 5 denoted the lowest confidence that the signal was present, and a rating of 8 indicated the highest confidence in their decision that the signal was present. After entering their confidence score, the subsequent trial would begin.

Before starting the experimental trials, participants completed 80 practice trials to familiarize themselves with the task. There were 20 microcalcification-present, 20 microcalcification-absent, 20 mass-present, and 20 mass-absent trials. The 80 DBT phantoms utilized here were from a different set of phantoms than the ones used in the experimental block. The practice trials followed the same procedure described above but with the addition of feedback at the end of each trial. The central slice was displayed at the end of signal-present trials (Figure 5.1.a, right). A fiducial marker was placed around the

signal to signify its (x, y) position. Text was also present above the image stimulus, denoting the central slice to inform participants where the signal was placed in the 3rd spatial dimension. If participants clicked on the trial, a circle was also overlaid on top of the image stimulus where they clicked so that they could visually correspond where they clicked with respect to the location of the signal. If the ground truth state was signal-absent, the text "ABSENT" was displayed on a gray background, and no image was presented. Participants had unlimited time to view the feedback.

2D search. Each image stimulus in this task corresponded to a single slice from one of the 160 DBT volumes used in the 3D search task. For the 80 signal-present stimuli, the central slices of the 80 DBT volumes were selected. The 32nd slice was chosen from the 80 signal-absent DBT phantoms for signal-absent stimuli.

We opted to use slices from the 3D DBT volumes for the 2D search task to isolate the effects of 2D versus 3D search on task performance. If we had chosen to use 2D mammogram phantoms instead, then the image statistics for this set of stimuli would differ from the image statistics of the cross-sectional slices in the 3D DBT phantoms (L. Chen et al., 2012), which could confound our analyses. This is mainly due to the image acquisition parameters and image reconstruction algorithms differing across 2D and 3D imaging modalities (Sechopoulos, 2013).

The trial procedure for the 2D search task, as shown in Figure 5.1.b, mirrored the steps for a trial in the 3D search task (e.g., specifying the signal to be looked for at the beginning, fixation cross to ensure proper eye tracker calibration, etc.). The main difference between the two tasks is the presentation of the image stimulus. For a 2D search trial, only a single image stimulus was presented to the participant, as shown in Figure 5.1.b, middle. Scrolling

was disabled, and no slice index tracker or custom scroll bar was present on the computer screen. All other aspects of the trial were consistent with a 3D search trial.

Participants completed 80 practice trials before beginning the 2D search experimental blocks. Again, we extracted 80 slices from the 80 DBT stimuli utilized in the 3D search practice blocks. Practice trials were broken up into eight 10-trial blocks.

Peripheral detectability task

Upon completion of the two experimental blocks described above, observers partook in a forced-fixation yes/no location-known-exactly detection task with the objective being to measure how well participants could detect each signal in their visual periphery (Lago, Sechopoulos, et al., 2020). Overall, there were 800 trials broken into two 2-hour sessions. All 800 stimuli were selected from a separate set of DBT phantoms than the ones used in the 2D/3D search tasks. The signal type was blocked in this task. Each block contained 400 trials. Half of the stimuli contained one kind of signal within a block, and the other half contained no signal. All stimuli were 2D slices of the 3D DBT phantoms. Signal-present stimuli corresponded to the signal's central slice within the DBT, like the 2D search task. Signal-absent stimuli were slices selected from the DBT phantoms that contained neither signal.

At the beginning of each trial, a black fixation cross and a black fiducial marker appeared on a gray background. The fixation cross was at the center of the screen, and the fiduciary marker was placed at a distance of 5 dva from the center of the fixation cross. The marker served as a visual cue that informed participants to covertly attend to that location before the stimulus would appear on the screen. The signal, if present, would always appear at that marked location. The five dva distance was chosen to prevent ceiling and floor effects

(i.e., perfect performance and chance performance, respectively). The marker could appear to the left or right, above or below the fixation cross. Placing the marker along the cardinal axes allowed us to account for anisotropies in visual processing across the vertical and horizontal meridians in the visual field (Abrams et al., 2012).

After staring at the fixation cross for 1 second, the image stimulus would appear on the screen for 200 ms. The fiducial marker was superimposed on the image stimulus, but the fixation cross was absent. Participants were instructed not to move their eyes. If they did, the trial would abort, and a "broken fixation" message would appear on the screen. (In this case, the trial would end, and participants would see the stimulus in a later trial). After viewing the stimulus, participants had to rate their confidence that the signal was present or absent at the cued location. We used the same rating scale as in the search tasks (i.e., ratings 1-8). Ground truth status was randomized across trials, and participants were informed that there was a 50% chance that the signal would be present at the cued location during any given trial.

Individual search performance measures and statistical analyses

Overview of dependent variables for 2D/3D search and peripheral detectability. We analyzed the area under the receiver operating curve (AUC), recognition errors, search errors, and the search area covered by the Useful Field of View (UFOV) to characterize observer performance across the 2D and 3D searches for both the microcalcification and mass signals (4 conditions). Lastly, we computed the AUC from the forced-fixation experiment to assess the peripheral detectability of each signal. Together, these analyses will provide the backdrop for explaining the expected benefits of the pooling algorithm in each of the four search conditions.

AUC. We first constructed a ROC curve using the rating scale (1-8) data for each participant in all four search conditions. We calculated the empirical area under the ROC curve using the trapezoidal method and then averaged AUCs across participants. Statistical significance was assessed for four pairwise comparisons: 2D-microcalcification vs. 2D-mass, 3D- microcalcification vs. 3D-mass, 2D-microcalcification vs. 3D-microcalcification, and 2D-mass vs. 3D-mass.

Recognition errors. Recognition and search errors are standard metrics for assessing human observer search performance in medical image perception tasks (Drew, Vo, Olwal, et al., 2013; Krupinski, 1996; Kundel et al., 1978; Lago, Jonnalagadda, et al., 2021). In our study, recognition errors occurred when a participant made a miss (rating < 5) and fixated on the signal during the trial at least once, regardless of the fixation duration. For the 2D search task, a *fixation on the signal* event occurred when the center of gaze position was at a distance less than or equal to 2.5 dva from the signal's centroid (x, y) coordinate. For the 3D search task, each signal spanned across multiple consecutive slices (i.e., they were visible on +- N slices from the central slice). Therefore, staring at the signal in 3D required the condition above and the concurrent slice where the fixation was recorded within N slices from the signal's central slice. For the microcalcification signal, N was set to 3. For the mass signal, N was set to 10. Recognition errors were tallied separately for each participant in the four conditions. The counts were divided by the total number of signal-present trials in a condition (40) to produce a recognition error rate (RER).

Search errors. Search errors were defined as the complement set of false negative responses. Participants did not fixate on the signal during the search and reported it absent. Search errors were also converted into rates (SER) by dividing the counts by 40. Together,

recognition errors, search errors, and hits summed to 40. Statistical significance for RER and SER were evaluated for the exact four pairwise comparisons as the AUC (8 comparisons in total across the two dependent variables).

UFOV coverage. The proportion of search area covered by the UFOV quantitatively measures how much participants explored with eye movements during the trial. This is an essential measure for this study because it allows us to ascertain whether human observers exhaustively scanned the search arrays, the second precondition outlined in the SDTmix framework.

For each recorded fixation in a trial, a circle with a radius of 2.5 dva was painted onto a binarized mask of a single slice of the DBT volume. Pixel values corresponding to the phantom tissue were converted to 1, and all background pixels were converted to 0. For the 3D search, if a participant fixated at one location and then proceeded to scroll through the slices, the circle was painted at that (x, y) location on each slice visited during the fixation. We calculated the union set of pixels covered by the UFOV and divided this count by the total number of pixels that comprised the binarized mask of either the 2D slice (2D search) or 3D volume (3D search). This procedure was done for all signal-present and all signal-absent trials, and the mean proportion of area covered was computed across all participants for each of the four conditions. Statistical significance was determined for all four pairwise comparisons discussed above.

Peripheral detectability. The other precondition in the SDTmix framework was a signal's peripheral detectability. Precisely, the signal's detectability needed to degrade rapidly as a function of retinal eccentricity.

The rating data from the 400 microcalcification and 400 mass trials per observer in the forced-fixation experiment were used to construct ROC curves. AUC was computed in the same manner as for the search task. Statistical significance was assessed between the two signals (1 pairwise comparison). One participant chose not to complete the mass peripheral detection block, and we omitted them from this analysis.

Statistical tests. We utilized a non-parametric bootstrap resampling procedure to obtain p-values for each pairwise comparison. First, we sampled, with replacement, stimuli (while maintaining 50% target-prevalence) and then participants. This was repeated 3,000 times. For each bootstrap iteration, we computed the mean AUC, mean RER, etc., for each of the four conditions. For a given dependent variable and pairwise comparison, we took difference scores across the 3,000 bootstraps. We counted the number of bootstrap difference scores more extreme than 0. We divided this count by 3,000 and multiplied the proportion by 2 to obtain a two-tailed p-value. All p-values were compared to an alpha level of 0.05.

Group decision-making performance measures and statistical analysis

Overview of the pooling algorithms. Three pooling algorithms were investigated in this work: average (AVG), majority vote (MAJ), and majority vote with exception (MAJe). Each model's performance for all four search tasks was evaluated for group sizes ranging from 2 to 9 members. The latter two algorithms were tested on groups of sizes 3, 5, 7, and 9, whereas the AVG algorithm was tested on all group sizes. For a given group size, m, there are $\binom{12}{m}$ possible combinations of groups. Considering one algorithm, one search task, and one group size, a model's expected performance was equal to the average performance across all possible combinations of groups. The descriptions of each pooling algorithm

below hone in on one search task and one group of a particular size for clarity, but the logic of the models can be generalized across groups, group sizes, and search tasks.

AVG. The AVG algorithm computed the mean rating across the members of the group. Each member's rating was weighted equally. We binarized the group's average rating to produce a yes/no decision. Specifically, the average rating was compared to a criterion—if the rating was above the criterion, the decision was yes. Otherwise, it was no. The criterion was set to the grand mean rating across all group members from all signal-present and signal-absent trials. Under an equal variance Gaussian model from Signal Detection Theory and 50% target prevalence, the midpoint between the noise and signal distribution means is the theoretically optimal criterion (Green & Swets, 1989). The grand mean discussed above is a point estimate of this midway point. The yes/no outcome was chosen as the decision variable instead of the average rating to be commensurate with the binarized decision variables inherent to the MAJ and MAJe algorithms.

MAJ and MAJe. The MAJ algorithm tallied the number of yes and no votes within the group for a given trial and selected the yes/no decision with the most votes. Each group member's rating was binarized (yes if rating > 4, no otherwise), and the majority decision was made. The MAJe algorithm was similar to the MAJ algorithm but with one crucial caveat. When the majority of group members voted no but at least one member produced a rating of 8 (i.e., the highest confidence that the signal was present), the MAJe decision variable was changed from no to yes. In all other instances, the MAJe decision variable was equal to the MAJ decision variable.

Pooling model performance metrics. Each pooling model's d' and proportion correct (PC) was benchmarked against two reference models: the mean observer (OBS) and a Signal

Detection Theory Independent model (SDT-IND). d'_{OBS} and PC_{OBS} were simply the average d' and average PC of the 12 participants. These point estimates did not change as a function of group size in our analyses and served to facilitate qualitative comparisons between a given algorithm and the average observer.

The SDT-IND model assumes that each group member's judgment is statistically independent and normally distributed and that the group's decision-making process is dominated by internal noise only. The shared noise generated from the image does not factor into the decision-making process. In other words, SDT-IND predicts the idealized benefit for aggregating group members' decisions—it is the ideal group (Juni & Eckstein, 2017). Below, we describe how d' and PC were calculated for AVG, MAJ, MAJe, OBS, and SDT-IND.

$\boldsymbol{d'}$. Sensitivity was calculated for the three pooling algorithms (d'_{AVG}, d'_{MAJ}, and d'_{MAJe}) using the standard formula, $d' = \Phi^{-1}(hit\ rate) - \Phi^{-1}(false\ alarm\ rate)$, where Φ^{-1} denotes the quantile function of the standard normal distribution. The hit rate is the proportion of signal-present trials where the group algorithm said yes, and the false alarm rate is the proportion of signal-absent trials where the group algorithm said yes. For a group of size m, the average hit rate and average false alarm rate were computed across the $\binom{12}{m}$ groups and plugged into the d' formula to obtain one sensitivity estimate per pooling model and group size. We assume a pooling algorithm's decision variable is sampled from the standard normal distribution on signal-absent trials. The decision variable is sampled from a unit variance normal distribution centered on d' for signal-present trials.

The sensitivity of the OBS was obtained by first transforming all 12 participant AUCs into $d'_a = \sqrt{2}\,\Phi^{-1}(AUC)$, an equivalent normal-normal sensitivity measure to d'. Here, we

assume the common binormal distribution where the pairs of transformed hit and false alarm rates form a straight line on the normal deviate axes (Metz, 1986; Swets, 1986). d'_{OBS} was the average d'_a across all 12 participants. The sensitivity for SDT-IND was computed in a 2-step process. First, for the j^{th} group of size m, $d'_{SDT-IND,j} = \sqrt{\sum_{i=1}^{m}(d'_{a,i})^2}$, where i denotes the i^{th} member of the group (Green & Swets, 1989; Sorkin et al., 2001). Next, we calculated the expected sensitivity of SDT-IND as follows: $d'_{SDT-IND} = \frac{1}{k}\sum_{j=1}^{k} d'_{SDT-IND,j}$, where k indicates the total number of unique groups of size m.

PC. PC_{AVG}, PC_{MAJ}, and PC_{MAJe} were each defined as the sum of hits and correct rejections divided by the total number of trials. Correct rejections occurred when a model said no on a signal-absent trial. We computed the mean PC across all groups of size m to obtain a point estimate for each model at that given group size. PC_{OBS} was defined as the mean PC across all 12 participants. For the SDT-IND PC prediction, $PC_{SDT-IND} = \frac{1}{k}\sum_{j=1}^{k} \Phi^{-1}(d'_{SDT-IND,j}/2)$, where k and j refer to the same variables as in the d' calculation and Φ^{-1} denotes the quantile function of the standard normal distribution. The right-hand side of the equation assumes an optimally placed criterion for an equal variance Gaussian SDT model with 50% target prevalence.

Statistical analysis for the pooling models. Our statistical analyses focused on two separate types of pairwise comparisons. The first comparison was made between the three algorithms within a given search task, totaling twelve comparisons across the four search tasks. The second type of comparison was made between two search tasks for a given algorithm, totaling twelve comparisons across the three algorithms. To facilitate these comparisons, we first computed their relative efficiency to SDT-IND. The efficiency is defined as the squared ratio of d's and this metric summarizes how each algorithm performs

with respect to a hypothetical reference group. The relative efficiency to SDT-IND, $\eta_{SDT-IND} = \left(\frac{d'_{pool}}{d'_{SDT-IND}}\right)^2$, where $d'_{pool} \in \{d'_{AVG}, d'_{MAJ}, d'_{MAJe}\}$, were computed separately for each group size and algorithm combination. In this work, $d'_{SDT-IND}$ can be thought of as d'_{ideal}. ${d'_{ideal}}^2$ is typically considered as being proportional to signal energy in low signal contrast detection tasks. Therefore, the optimal group detector can achieve the same performance as a particular pooling algorithm using only $100 * \eta_{SDT-IND}\%$ of the signal energy needed by the pooling algorithm under comparison (Sorkin et al., 2001).

To test for significant differences in relative efficiencies between 2 pooling algorithms or between two search conditions for one algorithm, we employed a similar bootstrap resampling procedure (i.e., sampling readers and cases with replacement) as discussed in the **Statistical tests** section above. However, we include one additional step for the pooling algorithms. For group sizes 5, 6, and 7, we sampled 500 random groups (without replacement) during each bootstrap iteration and computed the expected performance across those 500 groups. The total number of unique groups was less than 500 for all other group sizes, and all groups were used to compute expected performance. This process was repeated for each of the 3,000 bootstrap iterations.

For a given search condition (e.g., mass 3D search), we averaged the relative efficiencies across all common group sizes between 2 algorithms under comparison. For example, to determine if the relative efficiency of the AVG algorithm was higher than the relative efficiency of the MAJ algorithm, we computed the mean efficiency across group sizes 3, 5, 7, and 9 for each algorithm. Then, we computed the difference in mean relative efficiency. On the other hand, for a given algorithm (e.g., AVG), we computed the mean relative efficiency across all group sizes for that algorithm for two search conditions. For instance, in

comparing the relative efficiency of the AVG algorithm in 3D microcalcification search versus 2D microcalcification search, we first computed the mean efficiency across all group sizes ranging from 2-9 for each search condition. We took a difference score across these means.

Lastly, we computed the proportion of 3,000 bootstrapped difference scores more extreme than 0 and multiplied this by 2 to obtain a 2-tailed p-value for a given pairwise comparison. In total, 24 pairwise comparisons were evaluated. We applied an FDR correction (Benjamini & Hochberg, 1995) to an alpha level of 0.05 for twelve comparisons. We chose to correct for the 12 comparisons instead of 24 because each pooling algorithm was based on the same data, and comparisons across algorithms within a search condition are correlated. Thus, we opted to correct for 12 comparisons to preserve statistical power.

5.3.2. Results

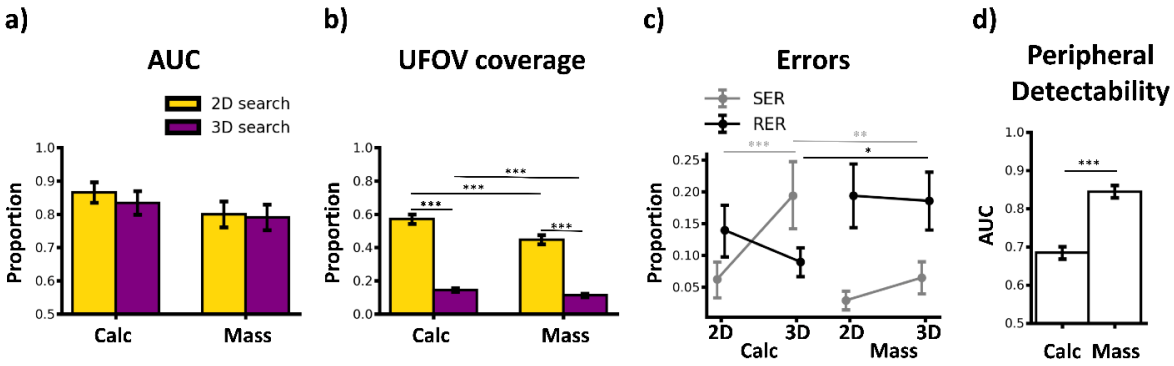

Figure 5.2. Search performance for the microcalcification (calc) and mass signals in the 2D and 3D search modalities. Microcalcification and mass peripheral detectibly are also shown. a) The mean AUC across all 12 observers for all four search conditions. The left set of columns refers to the mean AUC for the microcalcification signal, and the right set of columns denotes the mean AUC for the mass signal. Gold bars indicate 2D search and purple bars denote 3D search. b) The mean proportion of the search area covered by the Useful Field of View (UFOV). The same labeling for the four conditions used in a) is persevered here. c) The mean proportions of search (gray lines) and recognition errors (black lines). The left set of lines refers to search and recognition errors for the microcalcification signal across 2D (left) and 3D (right). The right set of lines refers to the same measures for the mass signal in 2D (left) and 3D (right). d) The mean AUC is the primary endpoint for assessing observer peripheral detectability for the microcalcification (left) and mass (right) signals. Error bars in each subplot denote 68% bootstrap confidence intervals (~ 1 standard error of the mean). Statistically significant differences are only shown for plotting clarity. * = $p < 0.05$, ** = $p < 0.01$, and *** = $p < 0.001$.

Characterizing 2D/3D search performance. Our first objective is to understand how observers performed in the 2D/3D searches for the two signals. Based on these analyses, we can ascertain whether a particular search task meets the two criteria outlined by the SDTmix framework—non-exhaustive coverage of the search array with eye movements and low peripheral signal detectability. These results will determine whether the predictions made by the SDTmix framework for the pooling algorithms apply to each search task. Below, we analyze the microcalcification performance across the two search modalities and then follow up with the same analyses for the mass signal. Lastly, we contrasted the performance of the two signals in 2D and then in 3D.

Microcalcification 2D versus 3D search. When observers looked for the microcalcification signal, there was no significant difference in overall performance when searching in 2D (AUC = 0.866) versus 3D (AUC = 0.834), $p = 0.293$ (Figure 5.2.a, left). However, the analysis of eye movement exploration (Figure 5.2.b, left) and gaze-contingent errors (Figure 5.2.c, left) suggest a more nuanced interpretation of the search performance across the two modalities. For instance, on average, observers covered 57% of the 2D DBT slices with their UFOV but only explored 15% of the 3D DBT volume, and this difference was statistically significant ($p < 0.001$). Moreover, the mean search error rate in 2D (0.063) was significantly lower than the mean search error rate in 3D (0.194), $p < 0.001$. However, the mean recognition error rate in 2D (0.14) was not significantly different from that in 3D (0.09), $p = 0.225$. Despite no significant differences in overall performance, observers under-explored the 3D DBT phantoms with eye movements, and this led to a substantial increase in search errors relative to the 2D search. These results suggest that at least the first

condition of the SDTmix framework was met for the 3D search of the microcalcification signal.

Mass 2D versus 3D search. Overall search performance for the mass signal in 2D versus 3D mirrored the relative difference in performance for the microcalcification signal (Figure 5.2.a, right). Specifically, the AUC in 2D (0.799) was not significantly different from the AUC in 3D (0.79), $p = 0.716$. Observers also covered markedly more of the search area with the UFOV in 2D (45%) than in 3D (11%), $p < 0.001$ (Figure 5.2.b, right). However, unlike the microcalcification signal, the mean search error rate in 2D for the mass (0.029) was not significantly different from that in 3D (0.065), $p = 0.09$. As reflected in Figure 5.2.c, right, the mean recognition error rate in 2D (0.194) was also not significantly different from that in 3D (0.186), $p = 0.807$. These results partially support the first criteria of the SDTmix framework for the mass 3D search. Still, the low search error rate suggests that observers sufficiently covered the 3D DBT phantoms with eye movements.

2D search microcalcification versus mass. The mean AUC in the 2D search for the microcalcification was higher but not significantly different from the mean AUC of the mass signal ($p = 0.171$). Observers, on average, also covered more of the DBT slices with the UFOV when tasked to look for the microcalcification signal than the mass signal ($p < 0.001$). This could result from how well observers can see the two signals in their visual periphery (Figure 5.2.d). The peripheral detectability of the mass (AUC= 0.844) was significantly higher than the peripheral detectability of the microcalcification (AUC = 0.68), $p < 0.001$. Not surprisingly, the microcalcification search error rate was higher than the mass search error rate in 2D. However, this observed difference was not significantly different from 0, $p = 0.179$. On the other hand, the microcalcification recognition error rate

was lower but not significantly different from the mass recognition error rate, $p = 0.367$. These results suggest that observers sufficiently explored the DBT slices when searching for the microcalcification. Although the signal is hard to detect in the visual periphery, the 2D search for the microcalcification fails to meet the first condition of the SDTmix framework. The 2D search for the mass signal also does not meet the requirements of the SDTmix model because the signal is readily detected in the visual periphery.

3D search microcalcification versus mass. Lastly, we compare the 3D searches for the microcalcification and mass signals. The difference in AUC was not significantly different from 0 for the 3D searches of the two signals, $p = 0.424$. Observers did cover a more significant proportion of the DBT volume with the UFOV when tasked to look for the microcalcification signal versus the mass signal, $p < 0.001$. However, observers made significantly more microcalcification search errors than mass search errors, $p = 0.007$. This makes sense, given that the microcalcification signal is more challenging to detect in the visual periphery than the mass signal. Conversely, observers made significantly fewer microcalcification recognition errors than mass recognition errors, $p = 0.036$. It is clear from this analysis that the microcalcification 3D search meets the two criteria outlined by the SDTmix model. Because the mass signal is more detectable in the visual periphery and the recognition errors are high, they sufficiently explored the 3D volume with eye movements. Therefore, the mass 3D search does not meet the criteria for the SDTmix model.

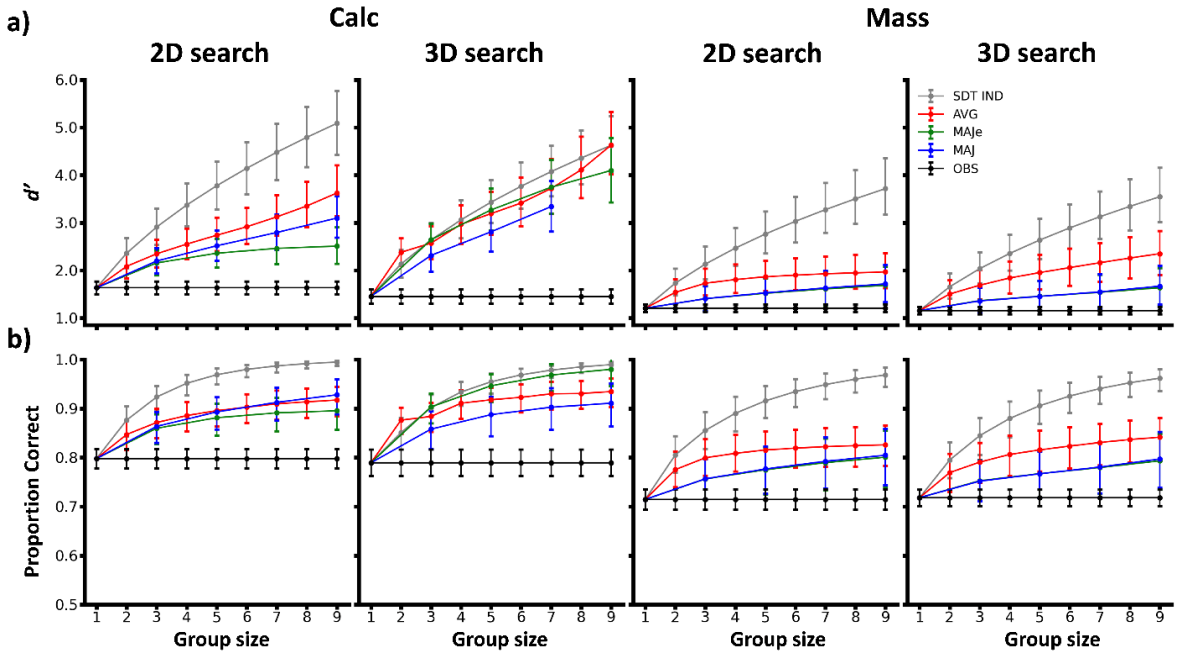

Figure 5.3. Pooling algorithms performance benchmarked against the mean observer (OBS) and SDT-IND. a) d' is plotted as a function of group size (1-9) for each of the four search conditions. The two subplots on the left refer to the 2D (left) and 3D (middle left) search conditions for the microcalcification signal. Similarly, the subplots on the right depict sensitivity as a function of group size for the mass signal in 2D (middle right) and 3D (right). OBS (black lines) refers to the mean d' across the 12 participants, and error bars represent the standard error of the mean. SDT-IND (gray lines) depicts predictions, as a function of group size, of the expected performance for a Signal Detection Theory model assuming independent judgments across participants. The three pooling algorithms (AVG-red, MAJe-green, and MAJ-blue lines) are plotted similarly. Note, in the microcalcification-3D search condition, d' is not shown for the MAJ model at a group size of 9 because the algorithm did not produce false alarms at that group size. As a result, d' could not be computed for that group size. b) The same performance analysis as a function of group size is shown as in a), but the dependent variable is proportion correct. In both a) and b), all error bars (except for OBS) represent 68% bootstrapped confidence intervals (~ 1 standard error of the mean). Additionally, all pooling models are anchored to OBS performance at group size 1.

Pooling model's performance across group sizes and search conditions. Figure 5.3. depicts the expected performance for all pooling algorithms as a function of group size. It includes OBS as a reference point, where OBS performance does not change as group size increases. The predictions made by SDT-IND demonstrate that expected performance (both d' and PC) increases as group size increases, which is consistent with previous work (Juni & Eckstein, 2017). This pattern holds across all four search conditions.

For the microcalcification signal, in both the 2D and 3D searches, the AVG and MAJ algorithms expected d' increase as group size increases (Figure 5.3.a, left and middle-left

subplots). For MAJe, the expected d' increases with group size for microcalcification 3D search but plateaus as group size increases for microcalcification 2D search. Moreover, the expected performance for the three algorithms in microcalcification 3D search aligns better with the predictions made by SDT-IND (i.e., more overlap in error bars) than for the microcalcification 2D search condition. All algorithms outperform the mean observer in the two microcalcification search conditions regarding d' and PC (Figure 5.3.a/b, left two subplots).

The results show a different pattern for the mass signal (Figure 5.3.a, middle-right and right subplots). The MAJ and MAJe algorithms deviate from the predictions made by SDT-IND for the mass 2D and mass 3D search conditions. As group size increases, the expected d' for the two algorithms increases but then plateau for larger group sizes. This is true for PCs as well. Interestingly, the expected d' increases more for the AVG algorithm relative to MAJ and MAJe as group size increases in 3D but not in 2D. Lastly, expected performance improvements for the two majority vote algorithms are incremental relative to OBS as group size increases.

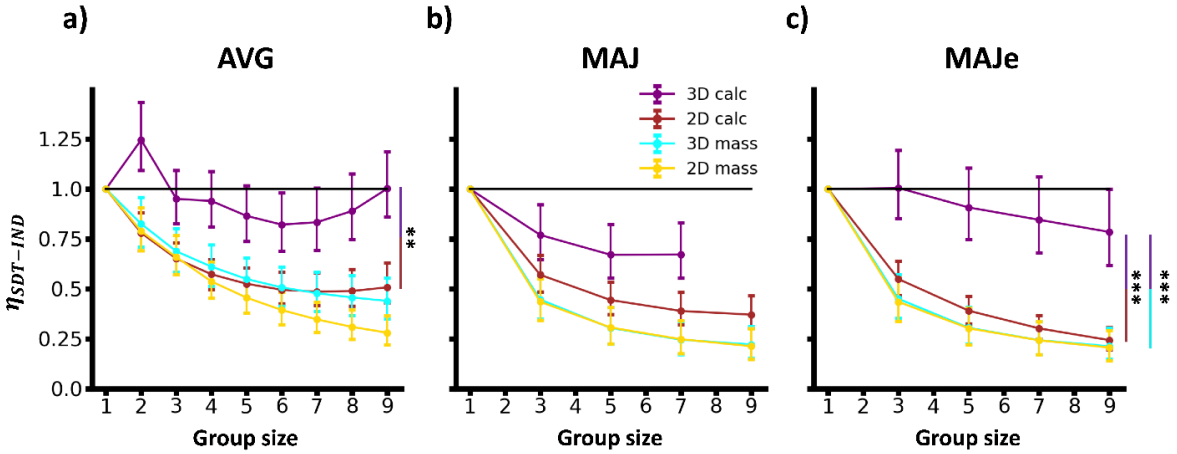

Figure 5.4. Relative efficiency to SDT-IND plotted as a function of group size for all four search tasks (colored lines) and stratified by pooling algorithm (subplots). The relative efficiency at a group size of 1 is anchored to 1 for all three subplots for display purposes only. The magenta lines refer to the efficiency of an algorithm in the 3D search for the microcalcification (calc) signal. The brown lines, cyan lines, and yellow lines denote the relative efficiencies of an algorithm in the microcalcification 2D search, mass 3D search,

and mass 2D search tasks, respectively. a) The relative efficiency for the AVG algorithm is plotted for all group sizes (2-9). b) The relative efficiency for the MAJ algorithm is plotted for all odd-numbered group sizes (3, 5, 7, and 9). Note that the relative efficiency at group size 9 for the microcalcification 3D search task is omitted because d' at group size 9 is not identifiable. c) The relative efficiency of the MAJe algorithm for all odd-numbered group sizes. The two unique colors that comprise a single vertical line on the righthand side of a given subplot correspond to a significant difference in mean relative efficiency, collapsed across all group sizes, between 2 search tasks. For example, the vertical bar on the right-hand side of the left subplot indicates that the mean relative efficiency of the AVG Algorithm in the microcalcification 3D search task is significantly higher than the mean relative efficiency of the AVG algorithm in the microcalcification 2D search task. ** = $p < 0.01$ and *** = $p < 0.001$. Error bars in all plots represent 68% bootstrap confidence intervals (~ 1 standard error of the mean).

Relative efficiency comparisons across search conditions for a given algorithm.

Each subplot in Figure 5.4. depicts an algorithm's relative efficiency to SDT-IND as a function of group size for all four search conditions. For the AVG algorithm (Figure 5.4.a), the mean relative efficiency in the microcalcification 3D search task ($\eta_{SDT-IND} = 0.943585$) was significantly higher than that in the microcalcification 2D search task ($\eta_{SDT-IND} = 0.563715$), $p = 0.009$. Similarly, the mean relative efficiency in the microcalcification 3D search task was substantially higher than that in the mass 3D search task ($\eta_{SDT-IND} = 0.569716$), $p = 0.030$, but did not survive an FDR correction. Neither the difference in mean relative efficiency between the mass 2D search versus the mass 3D search nor the difference in mean relative efficiency between the mass 2D search and the microcalcification 2D search reached statistical significance (see Table 5.1.).

Combining observer judgments using the MAJ pooling method (Figure 5.4.b) provided different expected benefits than the AVG model. In particular, the pairwise comparisons in the relative efficiency between search conditions were not significant (Table 5.1.). The biggest observed difference in mean relative efficiency was between the microcalcification 3D search ($\eta_{SDT-IND} = 0.704410$) and the mass 3D search ($\eta_{SDT-IND} = 0.332433$), $p = 0.017$. However, this comparison did not survive an FDR correction. On the other hand, the MAJe pooling model induced performance differences similar to those of the AVG pooling

method (Figure 5.4.c). For instance, the relative efficiency in the 3D microcalcification search condition ($\eta_{SDT-IND} = 0.885901$) was significantly higher than the relative efficiency in the 2D microcalcification search condition ($\eta_{SDT-IND} = 0.371643$), $p < 0.001$ and the 3D mass search condition ($\eta_{SDT-IND} = 0.303189$), $p < 0.001$. All other pairwise comparisons did not reach statistical significance (Table 5.1.).

Algorithm	**Search condition 1**	**Search condition 2**	**Mean $\eta_{SDT-IND}$ condition 1**	**Mean $\eta_{SDT-IND}$ condition 2**	**P-value**
AVG	Calc-2D	Mass-2D	0.563715	0.471633	0.344667
		Calc-3D		0.943585	**0.009333**
	Mass 3D	Mass 2D	0.569716	0.471633	0.390000
		Calc 3D		0.943585	0.030000
MAJ	Calc 2D	Mass 2D	0.468217	0.330291	0.194000
		Calc 3D		0.704410	0.134000
	Mass 3D	Mass 2D	0.332433	0.330291	0.977333
		Calc 3D		0.704410	0.017333
MAJe	Calc 2D	Mass 2D	0.371643	0.296951	0.418667
		Calc 3D		0.885901	**0.000667**
	Mass 3D	Mass 2D	0.303189	0.296951	0.943333
		Calc 3D		0.885901	**P < 3.33e^{-4}**

Table 5.1. Comparisons in mean relative efficiency (collapsed across all groups and group sizes) between search tasks for each pooling algorithm. Boldface p-values represent statistically significant differences that survived an FDR correction. We use the notation "Calc" to refer to the microcalcification signal.

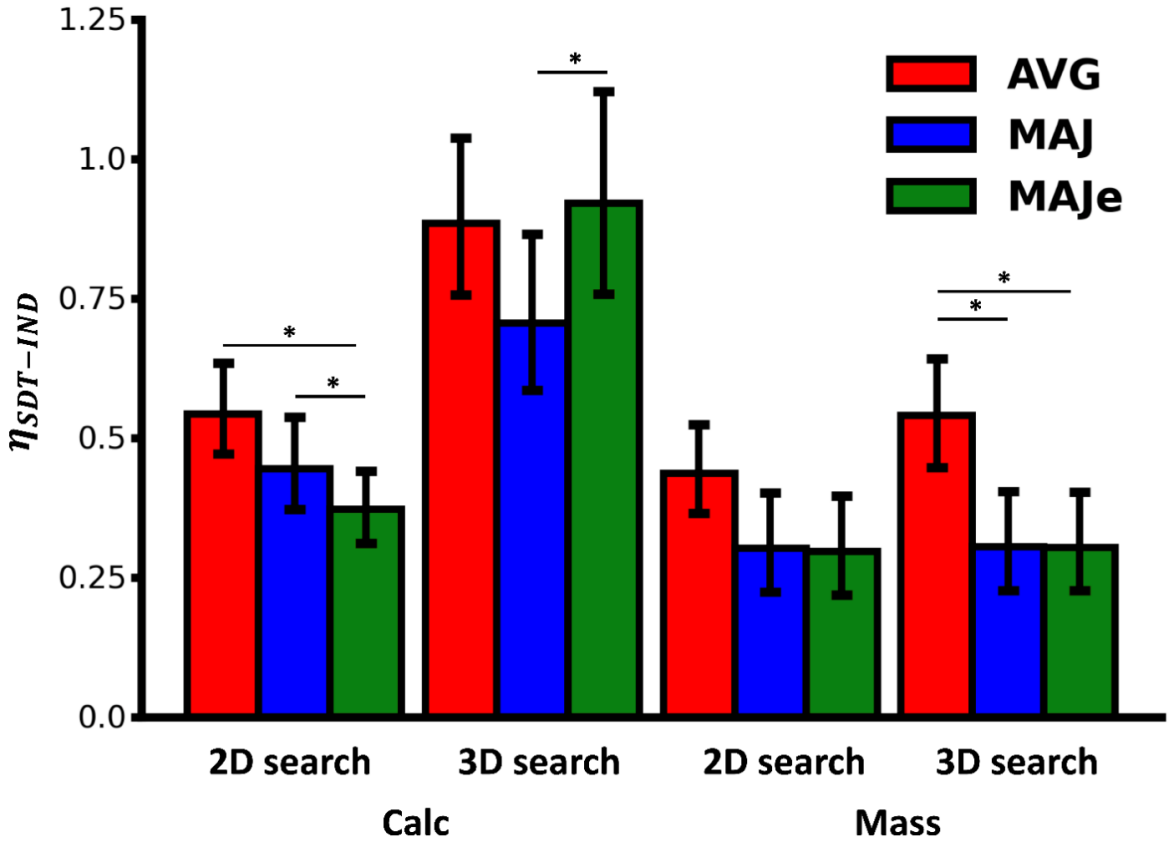

Figure 5.5. Mean relative efficiency collapsed across group size for each pooling algorithm stratified by search condition. From left to right, each cluster of bars denote the relative efficiencies of the three pooling models in the microcalcification (calc) 2D search, microcalcification 3D search, mass 2D search, and mass 3D search conditions. Red bars denote mean relative efficiencies for the AVG algorithm, blue bars denote the same metric for the MAJ algorithm, and green bars denote the efficiencies for the MAJe algorithm. All error bars represent 68% bootstrap confidence intervals (~ 1 standard error of the mean). * = $p < 0.05$.

Relative efficiency comparisons between algorithms for a given search condition.

Our last analysis focuses on contrasting the performance of the three algorithms within a single search condition. Figure 5.5. shows the mean relative efficiency, collapsed across group sizes, of each pooling algorithm within a given search condition. In the 2D search for the microcalcification signal (Figure 5.5., left), the mean difference in relative efficiency between the AVG and MAJe pooling methods was significantly different ($\Delta_\eta = 0.171196$, $p = 0.010667$). In contrast, the mean difference in relative efficiency between the AVG and MAJ algorithms was not significantly different from 0 ($\Delta_\eta = 0.098834$, $p = 0.153333$). The

mean difference in relative efficiency between the MAJ and MAJe pooling algorithms was also statistically significant ($\Delta_\eta = 0.072362$, *p = 0.04133*).

For the microcalcification 3D search condition (Figure 5.5., middle-left), The AVG pooling model's mean relative efficiency was not significantly higher than the MAJ pooling method ($\Delta_\eta = 0.178662$, *p = 0.05000*), nor was it considerably lower than the MAJe pooling model ($\Delta_\eta = 0.036465$, *p = 0.821333*). In comparing the two majority vote schemes, the MAJe algorithm's mean relative efficiency was significantly higher than the mean relative efficiency of the MAJ algorithm ($\Delta_\eta = 0.215128$, *p = 0.036667*).

How do the algorithms compare when observers were tasked to look for the mass signal? In the 2D search (Figure 5.5., middle-right), the AVG algorithm's relative efficiency was not significantly higher than the MAJe algorithm's relative efficiency ($\Delta_\eta = 0.138911$, *p = 0.105333*), nor was it significantly higher than the MAJ algorithm ($\Delta_\eta = 0.134782$, *p = 0.104000*). The mean relative efficiencies of the MAJ and MAJe algorithms were also not significantly different ($\Delta_\eta = 0.004130$, *p = 0.38533*).

Unlike the 2D search, in the 3D search for the mass signal (Figure 5.5., right), the mean relative efficiency of the AVG algorithm was significantly higher than the mean relative efficiency of the MAJe algorithm ($\Delta_\eta = 0.235838$, *p = 0.013333*). Similarly, the AVG algorithm had higher mean relative efficiency than the MAJ algorithm ($\Delta_\eta = 0.234185$, *p = 0.013333*). The difference in relative efficiencies across the two binary voting methods was not significantly different from one another ($\Delta_\eta = 0.001653$, *p = 0.843333*).

5.3.3. Discussion

At the outset, we posited that the 3D search for the small microcalcification signal meets the two preconditions spelled out in the SDTmix framework: low detectability of the signal

in the visual periphery and non-exhaustive search with eye movements in the 3D volumetric image. Indeed, our results show that observers had low peripheral detectability of the microcalcification signal in the visual periphery (Figure 5.2.d). Moreover, observers under-explored the 3D volumetric images, which is reflected by the fact that they explored significantly less in 3D than in 2D with the UFOV (Figure 5.2.b). Although there is no definitive threshold for under-exploration or non-exhaustive eye movement coverage, the results concerning the microcalcification search errors corroborate the claim that observers under-explored in 3D. Specifically, they made significantly more search errors in 3D than in 2D, and they made considerably more microcalcification search errors than mass search errors in 3D.

Given that the 3D search for the small microcalcification signal met the two preconditions of the SDTmix framework, does the performance of the group decision-making algorithms align with the predictions outlined under this framework? Specifically, do the AVG and MAJe pooling models, but not the MAJ model, have higher expected performance in the 3D search than the 2D search for the microcalcification signal? Additionally, do the AVG and MAJe pooling models outperform the MAJ pooling model in the 3D search for the microcalcification signal but not in the 2D search? Our results indicate that the relative efficiency of the AVG algorithm was significantly higher in the 3D search than in the 2D search (Figure 5.4.a). Similarly, the MAJe algorithm had a higher relative efficiency in the 3D search than in the 2D search (Figure 5.4.c), but the MAJ algorithm did not follow the same trend as the other two models (Figure 5.4.b). Regarding the second prediction, we observed that the MAJe but not the AVG pooling model had a significantly

higher relative efficiency than the MAJ pooling method in the 3D search for the microcalcification (Figure 5.5., middle-left).

We included an analysis of the group decision-making models for the 2D microcalcification search to highlight that the expected benefits are unique to the 3D search. Despite the signal being hard to detect in the visual periphery, it is relatively easy to visually scrutinize most regions of the 2D DBT slice with eye movements. Observers explored significantly more with the UFOV in 2D for the microcalcification signal than the mass signal (Figure 5.2.b). Thus, we would not expect the AVG or MAJe algorithm to have significantly higher relative efficiencies than the MAJ algorithm. Figure 5, left, showed that the relative efficiencies between the AVG and MAJ pooling models were not significantly different from one another. This finding aligns with a similar comparison between these two algorithms for a single-location detection task reported in (Juni & Eckstein, 2017), a detection task that did not meet the low peripheral detectability condition outlined by the SDTmix framework. Interestingly, the AVG and MAJ algorithms had significantly higher relative efficiencies than the MAJe pooling model. This discrepancy from our prediction can be explained by the fact that on signal-absent trials, at least one group member produced a false alarm with a rating of 8, the highest confidence in a signal-present decision. This is the only mechanism by which the relative efficiency of the MAJ pooling model can be higher than the MAJe pooling model, given that the two majority vote models differ in their decision if at least one person in the group produces a rating of 8.

Our inclusion of the search for the mass signal in 2D and in 3D and the subsequent analyses between search conditions and pooling models helped to juxtapose the unique benefits of group decision-making for the 3D search of the microcalcification signal. First,

the mass signal is more detectable in the visual periphery than the microcalcification signal (Figure 5.2.d). Since observers can see that signal better in the visual periphery, they would need to explore less with eye movements because their peripheral vision would compensate for the need to execute additional eye movements. We did find that observers explored less with their UFOV when searching for the mass than the microcalcification in both the 2D and 3D searches (Figure 5.2.b). We would not expect to see any difference in relative efficiency between the 2D and 3D searches for the mass signal for a given algorithm. Table 5.1. confirms that regardless of the pooling model under consideration, the difference in relative efficiency between the mass 2D search and 3D search did not reach statistical significance.

So how do the algorithms differ from one another, in terms of relative efficiency to SDT-IND, in the mass 2D search and the mass 3D search? The mass 2D search is directly opposite to the microcalcification 3D search concerning the two preconditions of the SDTmix framework. We expect the most negligible relative efficiency differences between the three pooling models. Figure 5.5., middle-left, confirms no significant differences in relative efficiency for all pairwise comparisons amongst the three pooling models. On the other hand, the analysis of the pooling models in the 3D search for the mass signal defied our predictions spelled out at the outset of this study because the AVG pooling model had a significantly higher relative efficiency to SDT-IND than the MAJ and MAJe pooling models. Sorkin et al. point out two possible reasons we observed this effect (Sorkin et al., 1998). First, the inefficiency of the Condorcet group's performance (i.e., majority vote models) is due to the least competent members' binary decision having an equal weight as the most competent group member's decision. Second, the graded information regarding the

signal's likelihood, encapsulated in the confidence score, is lost when the rating is converted to a binary decision.

The results of Experiment 1 provide a strong argument for applying the SDTmix framework to real-world visual search tasks involving analyzing complex image data. In the next experiment, we take a similar analytic approach as here but evaluate group decision-making with expert radiologists rather than trained human observers. Our goal is to determine whether the predictions outlined by the SDTmix extrapolate across domain expertise.

5.4. Experiment 2

5.4.1. Methods

Participants

Twelve radiologists (41% female, age range 27-35) participated in this study. Data were collected at the Radiological Society of North America conference in 2017.

Apparatus

The radiologist viewed stimuli on a medical-grade grayscale DICOM-calibrated monitor (5Mpx). They sat at a viewing distance of ~75 cm in a darkened room. Stimulus presentation, recording of mouse scroll movements, and all other aspects of the experiment were coded in Psychtoolbox (Kleiner et al., 2007).

Stimuli

The radiologists searched through 28 DBT phantoms in total. The 28 DBT phantoms were a subset of the 3D DBT phantoms seen by the undergraduate observers in Experiment 1. Each radiologist saw a random sample of seven microcalcification-present stimuli from a subset (14 stimuli) of 40 DBT images containing the microcalcification signal. Another

seven stimuli were randomly sampled from 14 of the 40 mass-present DBT phantoms shown to the undergraduate observers. Lastly, 14 DBT phantoms were sampled from a set of 28 unique signal-absent DBT phantoms seen by the observers in Experiment 1. Half of those stimuli were mass-absent, and the other half were microcalcification-absent.

Search task

The radiologists performed the same yes/no localization task (50% prevalence and signal-known-exactly) in 3D as the observers in Experiment 1. Each radiologist saw the 28 DBT phantoms in a randomized presentation order (i.e., randomized across all combinations of ground truth status and signal type). They completed the task in a single block. They rated their confidence in their decision on a scale of 1-4, with 1 representing the highest confidence in the signal's absence and 4 indicating the highest confidence in the signal's presence.

Grouping radiologist decisions together

Our analysis of the radiologist data is restricted to the 3D search of the mass and microcalcification signals. Furthermore, we assess the expected performance of the AVG, MAJ, and MAJe pooling models at group size three. We chose to evaluate a group size of three for two reasons. First, the randomized assignment of unique DBT phantoms to each radiologist created instances where all 12 radiologists did not see a single DBT phantom. For a signal-present stimulus, up to 8 of the 12 radiologists (mean number of radiologists = 6, std = 2) interacted with that DBT phantom. One radiologist was assigned to look for the mass signal for certain signal-absent stimuli, while another radiologist was asked to look for the microcalcification signal. Consequently, the total number of radiologists who saw a single mass-absent or microcalcification-absent DBT phantom varied up to 7, but, on

average, was 3 (std = 2). Second, independent double reading is employed in various European countries where two radiologists examine a single case (Geijer & Geijer, 2018). In the case of discordant opinions between the two readers, a third radiologist (or more) can be brought in for arbitration. Thus, our focus on groups of 3 emulates, to some extent, what is practiced in a real-world setting.

5.4.2. Results

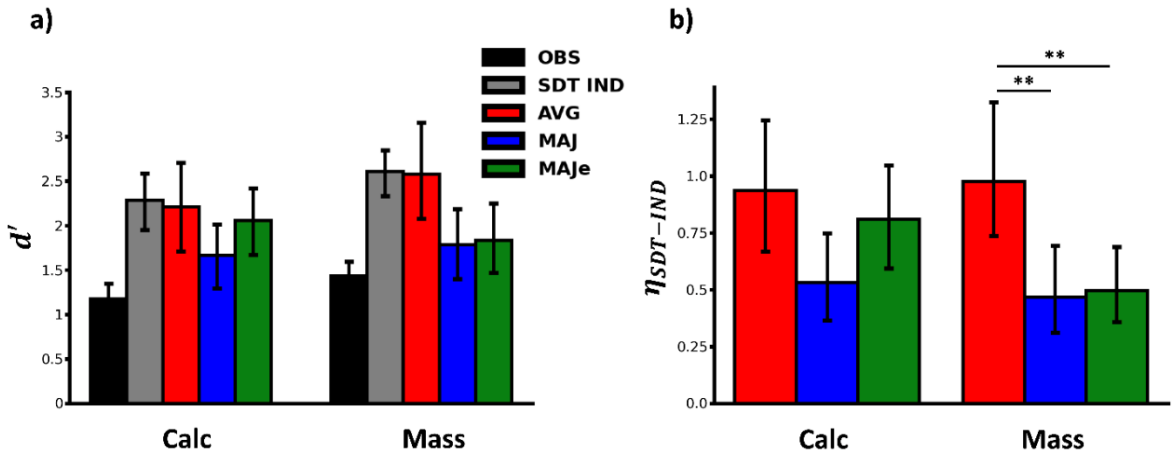

Figure 5.6. Pooling algorithms for a group size of three were applied to the radiologist's judgments in the 3D search for microcalcification and mass signals. a) d' is plotted for the microcalcification (calc) signal (left) and the mass signal (right) for the mean observer or OBS (black), the predictions made by SDT IND (gray), and the three pooling models (from left to right, AVG-red bars, MAJ-blue bars, MAJe-green bars). b) The relative efficiency to SDT IND is plotted for the three pooling models for the microcalcification signal (left) and the mass signal (right). The same color-coding scheme used to denote each algorithm in a) is used in b). The error bars for the mean observer represent the standard error of the mean. All other error bars denote 68% bootstrap confidence intervals (~ 1 standard error of the mean). ** = $p < 0.01$.

Our goal in Experiment 2 was to ascertain whether the expected benefits of the group decision-making models for complex visual search tasks extrapolate across expertise levels. Figure 6a, left, depicts the d' for the mean radiologist, the prediction made by SDT-IND, and the three pooling models for the 3D search of the microcalcification signal. Of note, all three pooling models have a higher d' than the mean radiologist, suggesting that there are performance benefits in this search task. Moreover, the AVG and MAJe pooling models

approximate the expected benefits predicted by SDT-IND. On the other hand, the MAJ pooling model performs the worst.

We computed the relative efficiencies of the three pooling models with respect to SDT-IND (Figure 5.6.b, left) to determine if the AVG and MAJe algorithms outperformed the MAJ pooling model on this search task. The AVG algorithm had the highest relative efficiency ($\eta_{SDT-IND} = 0.936833$). However, the relative efficiency was not significantly higher than that of the MAJe model ($\eta_{SDT-IND} = 0.809997$, $p = 0.547368$) nor that of the MAJ pooling model ($\eta_{SDT-IND} = 0.529757$, $p = 0.086737$). In comparing the relative efficiencies of the two majority vote models, we observed that the MAJe pooling scheme was not significantly higher than the MAJ pooling method ($\Delta_\eta = 0.280240$, $p = 0.053895$). Although the differences in relative efficiencies across the three pairwise comparisons did not reach statistical significance, the fact that the relative efficiencies of the AVG and MAJe models approached 1, whereas the MAJ relative efficiency resided near 0.5 suggests that the former two methods of aggregating radiologist decisions are effective at improving the 3D search of the microcalcification signal.

Does grouping radiologists' decisions together provide similar effects for the 3D search of the mass signal? Figure 5.6.a, right, shows that the three pooling models had higher sensitivity than the average radiologist. Similar to the microcalcification 3D search, the d' of the AVG model approximated what would be predicted by SDT-IND for the mass 3D search. However, the two majority vote models had sensitivities closer to the average radiologist than the prediction made by SDT-IND. Thus, the sensitivities of the three models with respect to the mean radiologist and SDT-IND are similar to the microcalcification 3D search sans the sensitivity of the MAJe model.

By focusing on the relative performance of the three pooling models with respect to SDT-IND, we note the relative efficiency of the AVG model ($\eta_{SDT-IND} = 0.976431$) was significantly higher than that of the MAJe model ($\eta_{SDT-IND} = 0.495526$), *p = 0.002526* and the MAJ model ($\eta_{SDT-IND} = 0.466941$), *p = 0.007579*. However, the latter two models were not significantly different in relative efficiency ($\Delta_\eta = 0.028584$, *p = 0.435368*). The slight difference in relative efficiency between the latter two pooling models suggests that radiologists were not highly confident when reporting "signal-present." On the other hand, when averaging the rating data from the groups of radiologists, their group performance improved markedly.

Lastly, we consider the difference in relative efficiencies across the two signals for a single algorithm. The AVG pooling method had near ceiling relative efficiency (i.e., $\eta_{SDT-IND} \sim 1$) for both the mass and microcalcification signals, and the difference was not statistically significant ($\Delta_\eta = 0.039598$, *p = 0.944000*). The MAJ pooling method also showed a negligible and nonsignificant difference in relative efficiency between the two signals ($\Delta_\eta = 0.062815$, *p = 0.764632*). On the other hand, the relative efficiency of the MAJe pooling model for the microcalcification signal was substantially higher than that for the mass signal ($\Delta_\eta = 0.314471$). However, this difference was not statistically significant (*p = 0.261053*). Our findings suggest that a simple averaging of confidence scores can improve radiologist group decision-making the most for 3D search, regardless of the signal being searched for. The MAJe pooling method performed well for the microcalcification signal but not the mass signal. On the other hand, the simple majority vote rule was the least efficient pooling method of the three, regardless of the signal radiologists were tasked to look for.

5.4.3. Discussion

Do the expected benefits of group decision-making generalize across expertise levels for complex tasks such as searching through large 3D volumetric medical images? In short, they mostly do. However, before assessing whether a subset of our predictions outlined at the outset hold across radiologists and trained observers, we briefly comment on differences in absolute performance across these two cohorts. On average, trained observers were slightly better at discriminating microcalcification-present trials from microcalcification-absent trials ($d' = 1.457272$) than radiologists ($d' = 1.173704$). However, for the mass signal, radiologists, on average, had a slightly higher sensitivity ($d' = 1.430509$) compared to the trained observers ($d' = 1.158341$).

Regarding the pooling models, we compare relative efficiencies to SDT-IND between radiologists and trained observers for a group size of 3. For the microcalcification signal, the relative efficiency of the AVG pooling model using the radiologist data was approximately 1% less than that using the trained observer's rating data. Similarly, the relative efficiencies for the MAJ and MAJe models utilizing the radiologists' decisions were 32% and 19% less than those using the trained observer's binary choices, respectively. On the other hand, for the mass signal, the relative efficiencies of the AVG, MAJ, and MAJe pooling models using the radiologist data were greater than those using the trained observer judgments by 42%, 5%, and 10%, respectively. In sum, the average trained observer and the corresponding groups of observers performed better when searching for the microcalcification signal. On the other hand, the radiologists and groups of radiologists performed better when searching for the mass signal.

Despite the difference in absolute performance across the two groups of observers in Experiments 1 and 2, our prediction regarding the expected performance benefits of the pooling models for the 3D search of the microcalcification signal was confirmed across expertise levels. A glance at Figure 5.5., middle left, and Figure 5.6.b, left reveals a common qualitative pattern among trained observers and radiologists regarding the differences in relative efficiencies between the three pooling models. In both groups, the AVG and MAJe pooling models have a higher relative efficiency to SDT-IND than the MAJ pooling model. The difference in relative efficiency between the AVG and MAJ pooling models for radiologists and trained observers was large and trending in the direction of statistical significance. On the other hand, the relative efficiency of the MAJe model was significantly higher than that of the MAJ for the trained observers but not the radiologists (although the difference in relative efficiency was trending towards significance). Why would the AVG and MAJe pooling models, but not the MAJ pooling model, have relative efficiencies close to 1?

Recall that SDT-IND makes predictions under the assumption of independent judgments across the member's decisions. Because the trained observers in Experiment 1 have limited peripheral detectability of the microcalcification signal and under-explored the 3D DBT phantoms with eye movements, the group member's idiosyncratic eye movement scan paths, on a trial-by-trial basis effectively produced an independent sampling of the large search space. Thus, it makes sense that the relative efficiencies of the AVG and MAJe are close to 1, suggesting that the two pooling models capture the independent judgments made by the group members. For the latter algorithm, if a majority of group members miss the microcalcification because they did not stare directly at it, but one member happens to, by

chance, fixate on the signal and produce the highest signal-present confidence score, then the MAJe pooling method is explicitly capturing the notion of independent sampling of the image data.

Prior work has outlined the search behavior of the radiologists used in this study (Lago, Jonnalagadda, et al., 2021). In particular, the radiologists underexplored the 3D DBT phantoms with eye movements to a similar extent as the observers in Experiment 1 (Figure 5.2.b). The under-exploration with eye movements degraded the radiologist's performance in the 3D search but not the 2D search for the microcalcification signal. The high relative efficiencies for the AVG and MAJe models suggest that radiologists also independently sampled the 3D data like the trained observers. Thus, the visual cognitive bottlenecks that arise from the foveated nature of the human visual system (Stewart et al., 2020; Tuten & Harmening, 2021) mediate the radiologist's performance to a greater extent than expertise for the 3D search of the microcalcification signal.

The pooling model performances from Experiment 2 mirror those in Experiment 1 concerning the 3D search of the mass signal. The AVG pooling method produced the highest relative efficiency to SDT-IND for trained observers and radiologists. In both instances, the relative efficiency of the AVG pooling method was significantly higher than that of the MAJe and MAJ pooling methods, as shown in Figures 5.5., right, and Figure 5.6.b, right, respectively. In Experiment 1, the difference in relative efficiencies between the MAJ and MAJe pooling models were marginal and nonsignificant (Figure 5.5., right), and this finding was actual for radiologists as well (Figure 5.6.b, right). Even though domain knowledge and expertise mediate observer performance (Nodine & Mello-Thoms, 2010; Wood, 1999) in difficult medical image perception tasks that are not limited by the

constraints of foveated vision (i.e., the mass is detectable in the visual periphery), we show that regardless of expertise level, the simple unweighted integration of confidence judgments is superior to a majority vote decision rule, or a variant of it.

Our last comparison between the two groups of observers focuses on the difference in relative efficiency between signals for a given pooling model. In Experiment 1, the MAJe pooling model had a significantly higher relative efficiency for the microcalcification 3D search condition than for the mass 3D search condition. In Experiment 2, this same difference in relative efficiency was large but did not reach statistical significance. One possible explanation for this discrepancy may be a matter of statistical power. We ran a nonparametric analog to an independent samples t-test for the radiologists because the same radiologists may have seen the microcalcification signal but not the mass signal. However, in Experiment 1, all observers saw the same images, which gave us more statistical power. For the AVG and MAJ pooling models, we observed no significant difference in relative efficiency between the two signals for the radiologists and the trained observers.

Our results confirm that pooling radiologists or trained observers' decisions together can induce the wisdom of crowds effect. Specifically, when searching for the microcalcification signal in 3D, the three pooling models tested here tend to outperform the average observer (Figure 5.3., middle-left) and the average radiologist (Figure 5.6.a, left). Moreover, the greatest benefit is applying a simple average over confidence estimates or a majority vote rule with an exception, as shown in Figure 5.5., middle-left, and Figure 5.6.b, left. Conversely, when searching for the mass signal in 3D, the unweighted average of group members' confidence scores outperforms the average observer and radiologist, as reflected

in Figures 5.3.a, right, and 5.6.a, right. It also provides a unique benefit not seen by the majority vote rules, as shown in Figures 5.5., right, and 5.6.b, right.

5.5. General discussion

Regarding overt search tasks, foveated vision plays a demonstrable role in mediating task performance (M. P. Eckstein, 2011; Najemnik & Geisler, 2005). This is especially true when the signal to be searched for is hard to detect in the visual periphery—but easy to detect when fixated upon—and the nature of the task, whether it be limited search time or a large search space, precludes exhaustive coverage of the image data with eye movements (Juni & Eckstein, 2017; Lago, Jonnalagadda, et al., 2021). One could consider these two factors as sufficient (but not necessary) conditions for generating independent processing of visual information between a group of observers looking at the same image stimulus. Group decision-making can exploit the independent sampling of visual information and the subsequent behavioral judgments (Sorkin & Dai, 1994) to induce the Wisdom of Crowds effect (Surowiecki, 2005).

In Experiment 1, observers searched through DBT phantoms for a microcalcification signal. The behavioral results replicated prior work, showing that observers under-explore 3D volumetric images with eye movements and miss small signals (i.e., high search error rate) because they are hard to see in the visual periphery (Lago, Jonnalagadda, et al., 2021). Thus, we predicted that the expected benefits of group decision-making would be at play in this search task because it meets the two conditions noted above. Indeed, we found that both averaging group members' confidence scores and an asymmetric maximum-confidence slating rule (i.e., only following the most confident signal-present decisions in the group) outperformed the mean observer to a great extent. These two pooling models also had higher

relative efficiencies to a statistically optimal pooling model (SDT-IND) than a commonly investigated majority vote rule. The latter finding suggests that the two pooling models can better utilize the independent information from each group member than the majority vote rule.

To demonstrate that the expected benefits of group decision-making were amplified and unique to the 3D search of the microcalcification signal, we had observers search for the same signal in a 2D DBT phantom slice. By design, this task makes it relatively easy to foveate most regions of the 2D slice in a time-efficient manner. Therefore, we would expect less independent sampling of visual information amongst observers than in the 3D search of the same signal. In comparing the performance of the pooling models across tasks, we observed that the averaging and majority vote with exception rules performed better in 3D than in 2D. These results helped bolster our claim that the benefits of pooling group member decisions together are enhanced for the 3D search of the microcalcification signal.

We included the 3D search for the mass signal in Experiment 1 to demonstrate that it is not just under-exploration that causes the enhanced benefit of group decision-making but also the signal's peripheral detectability. Again, we found that averaging confidence scores and the majority vote with exception pooling methods aligned better with SDT-IND in the 3D microcalcification search task than in the 3D mass search task. Together, these findings across both search modality (2D versus 3D) and signal type (microcalcification versus mass) clearly show the role of foveated vision inducing the benefits of group decision-making in complex perceptual tasks such as interpreting 3D medical images for detecting small signals.

A critical limitation of Experiment 1 is that we utilized trained undergraduate observers without prior experience interpreting and searching through medical images. Thus, it is

unclear if our results concerning the benefits of group decision-making extend across expertise levels. Although there is a large body of literature replicating the Wisdom of the Crowd effect with medical professionals (Brennan et al., 2019; Hasan et al., 2023; Kattan et al., 2016; Kurvers et al., 2016; Wolf et al., 2015) because of the known heterogeneity in task performance amongst radiologists (Beam et al., 2003), there has not been an investigation of the benefits of grouping radiologists decisions for screening tasks that involve large 3D volumetric images.

Experiment 2 tested whether the expected benefits of group decision-making amongst medical professionals extend beyond 2D displays. We found a strikingly similar pattern to Experiment 1 when assessing the relative performance benefits between the three pooling algorithms for the 3D searches of the microcalcification and mass signals. Like Experiment 1, we saw the performance advantage of the average and majority vote with exception rules over the majority vote rule for the 3D search of the microcalcification. We also identified that the average pooling method was superior to the two majority vote rules for the 3D search of the mass signal, following a similar pattern seen in Experiment 1. These similarities held despite differences in absolute performance across the two groups of subjects tested in experiments 1 and 2. The high relative efficiencies of the average and majority vote with exception rules for the 3D search of the microcalcification suggest that radiologists also sampled the 3D data independently when performing the visual search task. These findings indicate that pooling independent judgments can overcome universal visual-cognitive bottlenecks that limit a human observer's task performance.

As discussed above, the literature is replete with studies investigating the performance improvements of various pooling models concerning the average observer for a host of

visual perception tasks (Balsdon & Clifford, 2018; Juni & Eckstein, 2017; Kattan et al., 2016; Saha Roy et al., 2021). We consider how the findings from this study compare with other accounts on the benefits of group decision-making. Figure 5.7. provides a graphical depiction of the relative efficiency of the AVG pooling model (for a group size of 3) to the mean observer for five studies in addition to this one. Here, we replace the denominator of the efficiency calculation with the mean d' of the observers in the study rather than the d' predicted by SDT-IND. We chose this relative efficiency metric because the predictions made by SDT-IND were not included in most cases. We also report on the AVG pooling model's performance because it performed well in experiments 1 and 2 and is a commonly investigated pooling model in the literature.

Across the 12 tasks plotted in Figure 5.7., the AVG pooling model outperformed the mean observer, which is reflected by the relative efficiencies being above 1. We plotted the relative efficiencies in descending order to exemplify how the expected benefits of averaging confidence scores change across diverse tasks ranging from medical image perception to face perception. Importantly, we see that tasks where (1) the signal is hard to detect in the visual periphery and (2) the design of the experiment precludes observers from foveating all regions of the search space have the highest relative efficiencies (tasks with a * next to them). These types of tasks produce scenarios where a varying subset of group members, by chance, foveate the signal and produce high-confidence signal-present decisions that boost the group's performance by compensating for the remaining group members who only processed the signal in their visual periphery.

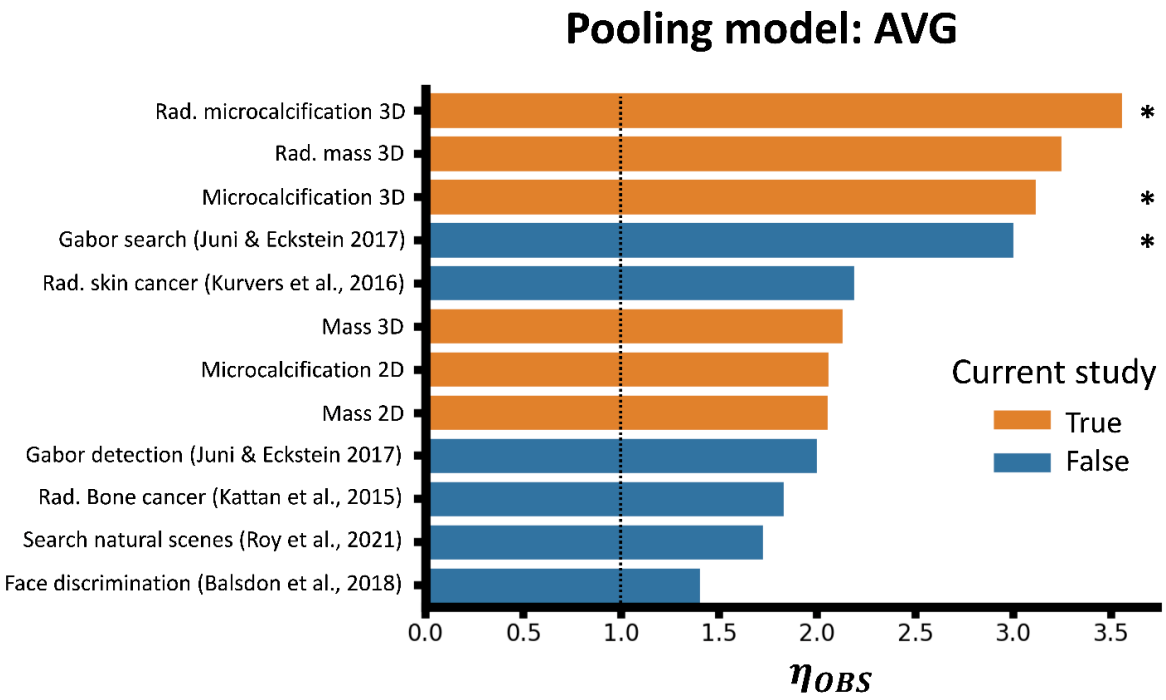

Figure 5.7. Comparisons of the relative efficiency of the AVG pooling model to the mean observer (OBS) across different tasks and studies. All relative efficiencies are reported for a group size of 3. Relative efficiencies (x-axis) are plotted across tasks (y-axis) in descending order. Orange bars denote tasks investigated in this study, and blue bars denote tasks reported in other studies in the literature. The vertical dotted line indicates a relative efficiency of 1 (i.e., no benefit in grouping decisions with respect to the average observer). Asterisks indicate tasks that strictly meet the two sufficient preconditions spelled out in the SDTmix theory framework to induce the benefits of group decision-making. Rad. = radiologist/dermatologist.

Our study has a few limitations that warrant future investigations into the benefits of collective intelligence for interpreting large 3D volumetric images. First, we did not modulate prevalence rates. In our study, observers knew at the outset of each trial that there would be a 50% chance that the signal would be present. However, manipulations to prevalence rates in search tasks have been shown to reduce search times, change an observer's criterion, and increase false negative decisions—see (Wolfe, 2012) for a review. Prevalence could negatively impact collective intelligence by increasing under-exploration with eye movements and thus increasing the correlation among group members' judgments. If all members under-explore because they are shortening their search time, then there is a higher probability that no observer in the group foveates the signal in 3D. Consequently,

members would not produce high-confidence signal-present decisions, which is a requirement for seeing the expected benefits of the AVG and MAJe pooling methods discussed in this work. Our analysis of 40 dermatologists looking for skin cancer (25% prevalence rate) showed that the AVG model is robust to prevalence effects in 2D images. Still, future work should assess whether this extends to 3D search scenarios.

Another limitation of our study concerns the satisfaction of search and the presence of multiple signals in a single medical image (Berbaum et al., 1990). Although recent work has shed light on this phenomenon in the context of 2D/3D search (Adamo et al., 2024), showing that 3D search reduces the likelihood of missing a second signal after finding the first, it is unknown how this phenomenon interacts with group decision-making. If the primary endpoint is recall and not localizing a lesion, then group decision-making may provide additional benefits when multiple signals are present. If two radiologists localize different signals because they visually scrutinized different regions of the same 3D image, then the group has successfully flagged potential cancer, and additional workup can be done to identify both signals.

VI. Conclusion

The purpose of this thesis is twofold. The first aim is to understand better why human observers under-explore 3D volumetric images with eye movements (Lago, Jonnalagadda, et al., 2021). By leaving much of the 3D image data "unexplored," observers tend to miss small signals that are hard to detect away from the point of fixation or near the fovea. They are highly likely to process the signal solely in their visual periphery during the search and report it as absent (search errors).

The second aim of this thesis is to investigate three methods for mitigating 3D search errors. The first method concerns a 2D synthetic view of the 3D data, an adjunct to interpreting the 3D volumetric image. The 2D synthetic view supplants a 2D mammogram, thus reducing the radiation exposure to the patient. The second method is a deep learning-based computer-aided detection tool that works in tandem with human observers as they perform the 3D search. The last method focuses on pooling independent judgments across a group of observers to harness the wisdom of crowds. Each approach has been deployed in practice within a clinical setting (Aujero et al., 2017; Lamb et al., 2022; Taylor-Phillips & Stinton, 2019a). However, the theoretical underpinning for how these methods can improve medical image perception by reducing 3D search errors is poorly understood. Therefore, the second aim of this thesis is to fill in these theoretical gaps.

At the outset, I proposed the following questions. (Q.1) When observers under-explore 3D images, what evidence do they base their quitting decision on? (Q.2) How do the 2D synthetic image and CADe influence observers' search strategies in 3D? (Q.3) Are there unique benefits gained from aggregating observers' judgments in 3D that are not seen in 2D searches? (Q.4) For each of the three methods (2D synthetic view, CADe, or group

decisions), which types of signals benefit the most? (Q.5) Lastly, which of these techniques is best suited to mitigate 3D search errors? Below I provide an answer (A.1, A.2, etc..) to each of these questions to further the field in having a better theoretical grasp of how to improve 3D medical image perception.

(A.1). Perceived peripheral detectability of targets may mediate how much people explore in 3D medical images. Chapter II directly investigates the interaction between 3D under-exploration and the search-termination criterion. It provides two significant contributions to this line of inquiry. First, it builds upon the paucity of empirical work theorizing about the specific evidence (i.e., quitting signal) humans consider when asking themselves, is it time to abandon the current search and move on to the subsequent trial? One prominent model for this in 2D displays posits that a quitting signal follows a drift-diffusion process (Ratcliff, 1978), and observers end their search once the signal exceeds a threshold (Wolfe & Van Wert, 2010). Factors like target prevalence can be incorporated into the model by adjusting the diffusion threshold (Wolfe, 2012). However, this approach treats the quitting signal as an internal variable to the searcher (a free fitting parameter) without explicitly linking the stopping criterion to the properties of the target and its relationship with the background image statistics or the size of the search display.

Chapter II posits that the quitting signal is proportionally related to the percentage of image area covered with eye movements when the target prevalence rate is set to 50%. The results from a series of experiments showed that humans utilize their perception of how well they can see various signals in their visual periphery to gauge how much search space they have covered. Once their internal estimate of the area explored with eye movements exceeded a stopping criterion, they terminated their search. This hypothesis produced

consistent exploration behavior between a small and large target with different peripheral detectability despite differences in reaction time, number of eye movements, and performance between the searches of the two signals. Importantly, this hypothesis is amenable to linking the peripheral detectability of various signals in arbitrary 2D and 3D backgrounds (i.e., not only items on a uniform background) to a stopping criterion, which is particularly pertinent for computational models of eye movements as they are often in need of a principled stopping criterion (Akbas & Eckstein, 2017; Hoppe & Rothkopf, 2019; Lago, Abbey, et al., 2020; Najemnik & Geisler, 2005; Zelinsky, 2008). The findings do not negate previous factors that influence the stopping criterion, which can influence the adopted threshold related to the perceived % area covered.

Chapter II's second contribution provides one possible explanation for why humans under-explore 3D volumetric images. The results from Chapter II Experiment 2 showed that humans tend to explore the 2D image plane of the 3D volume to the same extent as an analogous 2D search task, suggesting that they may be under-exploring the 3D volume because they are defaulting to sufficiently covering the 2D image plane without regard for areas of the 3D volume that have yet to be explored. This finding speaks more to how 3D images are displayed to humans when performing visual searches rather than the features that comprise the image. Tatler showed that the simple presentation of an image on a computer monitor screen induced the central fixation bias (Tatler, 2007), which fundamentally changed how the field evaluated contemporary saliency models that predicted fixation distributions in natural scene-viewing tasks (Bylinskii et al., 2019). In a similar vein, the presentation of a series of cross-sectional slices viewed one at a time on a computer monitor may interact with human observers' entrenched oculomotor strategies of viewing

single images on computer monitors, and this interaction may cause them to under-explore the 3D volumetric image.

(A.2). Search aids guide eye movements to areas in 3D that would have otherwise not been processed with foveal vision. Chapter III evaluated how a 2D synthetic view (2D-S), used as an adjunct to interpreting a 3D volumetric image, mitigated 3D search errors. The 2D-S reduced search errors for a small but not a large signal. By applying a dimensionality reduction to the image data, the 2D-S markedly reduced the spatial location uncertainty of the small signal that was hard to detect in the visual periphery. Human observers were able to scrutinize the 2D-S and mark suspicious locations efficiently. The marked locations helped guide eye movements in 3D, reducing search errors. To demonstrate the influence of the 2D-S on the 3D search, an image computable Foveated Search Model (FSM) was employed. The FSM explicitly models the effects of foveated vision on peripheral signal detectability. It took each human observer's unique 2D and 3D scan paths as input, along with the image data, and produced a test statistic representing the presence/absence of a signal. The model performance correlated well with the human observer performance. It captured the relative differences in AUC across the 2D-S search, 3D search, and 2D-S + 3D search. Together, these results showed how a 2D-S can mitigate the detrimental impact of under-exploration of 3D volumetric images.

One interesting takeaway from the findings of Chapters II and III is how the 2D-S serves as a natural complement to the oculomotor strategy of focusing on the 2D image plane during the 3D search. Since the FDA approved digital breast tomosynthesis (DBT) for early breast cancer detection in 2011, all interpretations of 3D DBT images require either a 2D mammogram or a 2D-S image (FDA approval in 2014) generated from the corresponding

3D DBT image. This thesis provides a theoretical lens for how a 2D-S adjunct can accommodate 3D search strategies.

Chapter IV investigates how a convolutional neural network (CNN)-based computer-aided detection (CADe) system helps human observers find small microcalcification-like signals in 3D DBT phantoms. This study demonstrated that the CNN-CADe was an effective tool for improving the 3D search of a small microcalcification signal by reducing 3D search errors. In particular, the cue boxes produced by the CNN-CADe, superimposed on top of the 3D image stimulus, guided human eye movements to suspicious locations that the fovea would have otherwise not processed.

An important finding from this study was that the 3D search benefits of the CNN-CADe were unique to the microcalcification-like signal but not a second larger mass-like signal. Moreover, an examination of individual differences showed a strong negative linear relationship between the proportion of area explored in 3D and the additional improvement in performance when using the CNN-CADe output during the search—those who under-explored the most in 3D benefitted when the aid was available.

The latter finding could serve as an additional data point in radiology residency programs for emphasizing the benefits and pitfalls of relying on an AI aid to interpret 3D volumetric medical images. If a radiologist tends to under-explore and rely heavily on the AI-CAD output, their sensitivity and specificity will be similar to that of the AI (Deza et al., 2019), at least for detecting small microcalcification signals

(A.3). Crowd wisdom and the corresponding idiosyncratic 3D scan paths are an effective method for reducing 3D search errors. Chapter V took a conceptually different approach to mitigating the detrimental impact of 3D under-exploration. Unlike Chapters III

and IV, which augmented the search process, Chapter V investigated the benefits of pooling independent judgments from multiple visual searchers interacting with DBT phantoms and 2D slices of DBT phantoms. The results indicated that either averaging independent confidence ratings or taking a majority vote with an exception rule (the exception being following the most confident target-present response in the group) but not a simple majority vote rule produced the greatest benefit for the 3D of the microcalcification signal. Specifically, the two models outperformed the average searcher. The models also induced a higher benefit here than in an analogous 2D search task for a microcalcification signal and a 3D search task for the large mass signal.

The findings were partially replicated with radiologists. However, more data needs to be collected to corroborate this. The application of radiologist crowd wisdom and breast cancer detection in 3D is particularly interesting because most countries that require double reading (e.g., Sweden, Norway, Australia, New Zealand, etc.) employ different rules for requiring a patient to undergo additional workup (recall). Some follow a majority vote, and others require just one of two radiologists to ask for a recall. Moreover, most of these countries still rely on 2D Full Field Digital Mammography. Our work and theoretical modeling suggest that if 3D DBT is adopted in these countries, it might affect the rules governing how to best combine radiologist judgments to form a consensus opinion.

Signal type	Imaging modality	Method for improving search		
		2D-S (Chapter III)	CNN-CADe (Chapter IV)	Wisdom of Crowds (Chapter V)
Small	2D	-	1.43	1.67
	3D	**2.25**	**2.16**	**2.53**
Large	2D	-	1.65	1.49
	3D	1.66	1.14	1.39

Table 6.1. Comparing relative efficiency across search tasks for a given method. The relative efficiency here is defined as: d' numerator = detectability of observers + search aid (or wisdom of crowds) and d'

denominator = average detectability of group on search task (without aid or wisdom of crowds). Boldface numbers represent the highest relative efficiency within a column (method).

(A.4). 3D search for the small signal benefits the most for each of the three methods investigated. Table 6.1. provides a summary of each method's added benefit to the 2D/3D search for the small and large signals. In particular, we computed the relative efficiency to the mean observer, the d' when searching with the aid (or pooling observers' judgments together), divided by the average d' without the aid (or no pooling across judgments). It is clear from the table that regardless of whether observers searched with the 2D-S, the CNN-CADe, or a majority vote with exception rule was taken across group members' decisions, the 3D search for the small signal benefited the most. Additionally, suppose we disregard the imaging modality factor (i.e., average the 2D/3D search relative efficiencies). In that case, we note that the small signal benefited more from the three methods than the large signal. In sum, each of these methods strongly interacts with the negative effect of under-exploration and the foveated nature of the human visual system.

Future work can look at methods for reducing recognition errors of larger signals in both 2D and 3D modalities that are related to visual masking and the human observer's inability to see through image and background noise.

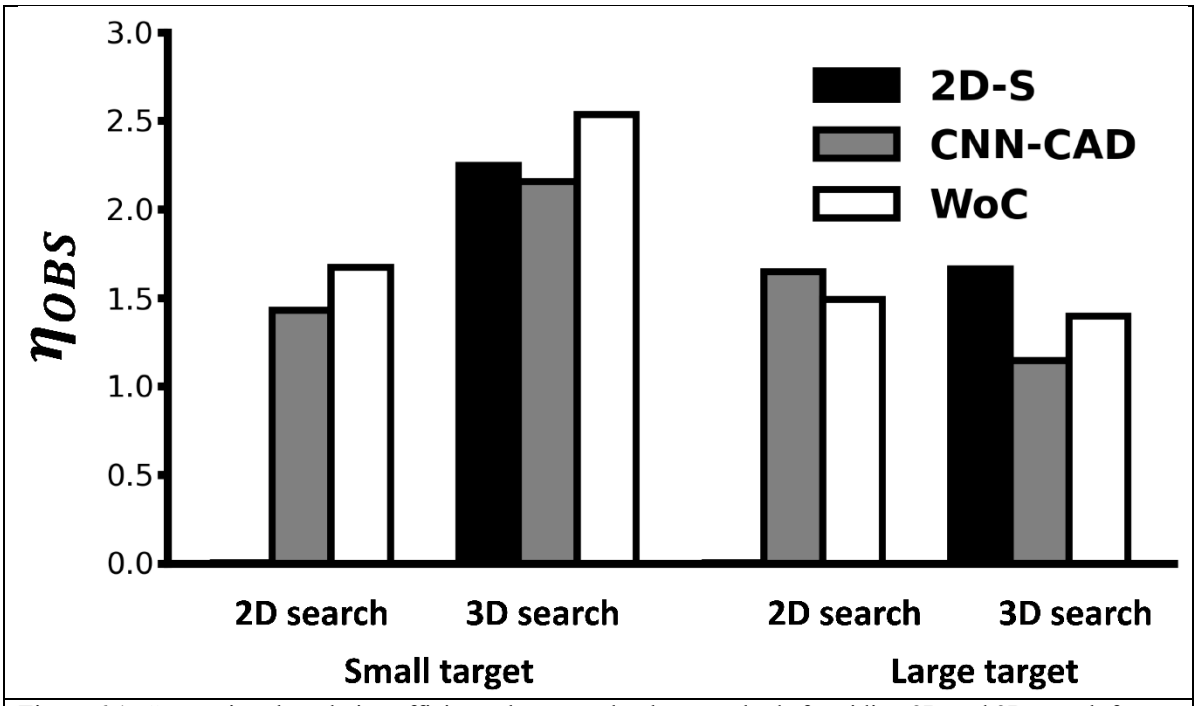

Figure 6.1. Comparing the relative efficiency between the three methods for aiding 2D and 3D search for small and large signals. From left to right: 2D search small signal, 3D search small signal, 2D search large signal, 3D search large signal. Black bars denote the 2D-S method, Gray bars denote the CNN-CADe method, and white bars represent the wisdom of crowds (majority vote with exception). For Chapter III, we exclude the calculation of relative efficiency for 2D since observers only interacted with the 2D-S.

(A.5). Comparing the three methods for mitigating 3D search errors. Figure 6.1 plots the relative efficiency of each method in Table 6.1 but stratified by search condition on the x-axis. Of interest, the wisdom of crowds (using the majority vote with exception rule) slightly outperforms the other 2 methods. Returning to A.3 for a moment, double reading with DBT may prove to be effective at improving the detection of small microcalcifications or microcalcification clusters. It would be interesting to see the results from future pilot studies using actual radiologists.

However, caution should be taken in emphasizing these comparisons across methods because the investigation of the 2D-S adjunct was done in correlated Gaussian noise and with signal uncertainty. On the other hand, the CNN-CADe method and wisdom of crowds analyses were investigated using DBT phantoms and no signal uncertainty. Additionally, the

results will depend on the overall accuracy achieved by the CNN-CADe relative to humans. For our case, the CNN-CADe correctly localized the signals in the DBT phantoms and single slices of the phantoms in 77.5% of the signal-present trials. On the other hand, when observers searched for the small signals in 2D and 3D, they localized the signals on 74.8% (sd=12.2) and 64.8% (sd=15.0) of signal-present trials, respectively. For the large signal 2D and 3D searches, they localized the signal on 68.1% (sd=12.8) and 60.4% (sd=8.4) of the signal-present trials. Thus, we would predict that if the CNN-CADe could localize the signals on more signal-present trials, then the CNN-CADe may have a higher relative efficiency than the wisdom of crowds method.

Concluding remarks. In conclusion, this thesis has increased our understanding of the mechanisms mediating human under-exploration of 3D volumetric images and the associated search errors. It has assessed potential solutions spanning from multiple readers, AI aids, and using 2D synthetic images as an adjunct to the 3D search for mitigating 3D search errors. Although there was an emphasis on Digital Breast Tomosynthesis, the results and findings are likely to be applicable to all scenarios in which human observers search through large image stacks, including other medical imaging modalities (e.g., CT, MR, or PET) and potentially other applications like CT scans of luggage in airports and satellite imagery.

VII. References

Abbey, C. K., & Barrett, H. H. (2001). Human-and model-observer performance in ramp-spectrum noise: Effects of regularization and object variability. *JOSA A*, *18*(3), 473–488.

Abbey, C. K., & Bochud, F. O. (2000). *Modeling Visual Detection Tasks in Correlated Image Noise with Linear Model Observers*. *PM79*, 629–655. https://doi.org/10.1117/3.832716.ch11

Abbey, C. K., & Eckstein, M. P. (2007). Classification images for simple detection and discrimination tasks in correlated noise. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *24*(12), B110-124.

Abbey, C. K., & Eckstein, M. P. (2014). Observer efficiency in free-localization tasks with correlated noise. *Frontiers in Psychology*, *5*, 345. https://doi.org/10.3389/fpsyg.2014.00345

Abbey, C. K., Samuelson, F. W., Zeng, R., Boone, J. M., Eckstein, M. P., & Myers, K. (2018). Classification images for localization performance in ramp-spectrum noise. *Medical Physics*, *45*(5), 1970–1984. https://doi.org/10.1002/mp.12857

Abrams, J., Nizam, A., & Carrasco, M. (2012). Isoeccentric locations are not equivalent: The extent of the vertical meridian asymmetry. *Vision Research*, *52*(1), 70–78. https://doi.org/10.1016/j.visres.2011.10.016

Adamo, S. H., Brem, R., & Mitroff, S. R. (2024). Comparing multiple-target search performance and the satisfaction of search effect between 2D and segmented-3D search. *Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment*, *12929*, 39–44. https://doi.org/10.1117/12.3005601

Adamo, S. H., Gereke, B. J., Shomstein, S., & Schmidt, J. (2021). From "satisfaction of search" to "subsequent search misses": A review of multiple-target search errors across radiology and cognitive science. *Cognitive Research: Principles and Implications*, *6*(1), 59. https://doi.org/10.1186/s41235-021-00318-w

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*(4), 045501.

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLOS Computational Biology*, *13*(10), e1005743. https://doi.org/10.1371/journal.pcbi.1005743

Alabousi, M., Zha, N., Salameh, J.-P., Samoilov, L., Sharifabadi, A. D., Pozdnyakov, A., Sadeghirad, B., Freitas, V., McInnes, M. D. F., & Alabousi, A. (2020). Digital breast tomosynthesis for breast cancer detection: A diagnostic test accuracy systematic review and meta-analysis. *European Radiology*, *30*(4), 2058–2071. https://doi.org/10.1007/s00330-019-06549-2

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, *11*(8), 909–918. https://doi.org/10.1016/j.acra.2004.05.012

Aujero, M. P., Gavenonis, S. C., Benjamin, R., Zhang, Z., & Holt, J. S. (2017). Clinical Performance of Synthesized Two-dimensional Mammography Combined with Tomosynthesis in a Large Screening Population. *Radiology*, *283*(1), 70–76. https://doi.org/10.1148/radiol.2017162674

Ba, A., Shams, M., Schmidt, S., Eckstein, M. P., Verdun, F. R., & Bochud, F. O. (2020). Search of low-contrast liver lesions in abdominal CT: The importance of scrolling behavior. *Journal of Medical Imaging*, *7*(4), 045501. https://doi.org/10.1117/1.JMI.7.4.045501

Badano, A., Graff, C. G., Badal, A., Sharma, D., Zeng, R., Samuelson, F. W., Glickman, S., & Myers, K. J. (2018). Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial. *JAMA Netw Open*, *1(7):e185474*.

Baker, J. A., & Lo, J. Y. (2011). Breast Tomosynthesis: State-of-the-Art and Review of the Literature. *Academic Radiology*, *18*(10), 1298–1310. https://doi.org/10.1016/j.acra.2011.06.011

Bakic, P. R., Barufaldi, B., Higginbotham D., S. P., Weinstein, K, E., Xthona, A., Kimpe, T., & Maidment, A. D. (2018). Virtual Clinical Trial of Lesion Detection in Digital Mammography and Digital Breast Tomosynthesis. *Proc.SPIE*, *10573*.

Ball, K. K., Beard, B. L., Roenker, D. L., Miller, R. L., & Griggs, D. S. (1988). Age and visual search: Expanding the useful field of view. *Journal of the Optical Society of America. A, Optics and Image Science*, *5*(12), 2210–2219.

Balsdon, T., & Clifford, C. (2018). Task Dependent Effects of Head Orientation on Perceived Gaze Direction. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02491

Banks, M. S., Sekuler, A. B., & Anderson, S. J. (1991). Peripheral spatial vision: Limits imposed by optics, photoreceptors, and receptor pooling. *JOSA A*, *8*(11), 1775–1787. https://doi.org/10.1364/JOSAA.8.001775

Barrett, H. H. (1990). Objective assessment of image quality: Effects of quantum noise and object variability. *JOSA A*, *7*(7), 1266–1278.

Barrett, H. H., & Myers, K. J. (2013). *Foundations of image science*. John Wiley & Sons.

Barrett, H. H., Yao, J., Rolland, J. P., & Myers, K. J. (1993). Model observers for assessment of image quality. *Proceedings of the National Academy of Sciences*, *90*(21), 9758–9765.

Barufaldi, B., Higginbotham, D., Bakic, P. R., & Maidment, A. D. A. (2018). OpenVCT: a GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis. In J. Y. Lo, T. G. Schmidt, & G.-H. Chen (Eds.), *Medical Imaging 2018: Physics of Medical Imaging* (Vol. 10573, pp. 1333–1340). SPIE. https://doi.org/10.1117/12.2294935

Beam, C. A., Conant, E. F., & Sickles, E. A. (2003). Association of Volume and Volume-Independent Factors With Accuracy in Screening Mammogram Interpretation. *J. Natl. Cancer Inst.*, *95*(4), 282–290. https://doi.org/10.1093/jnci/95.4.282

Beck, M. R., Peterson, M. S., & Vomela, M. (2006). Memory for where, but not what, is used during visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 235–250. https://doi.org/10.1037/0096-1523.32.2.235

Becker, M. W., Rodriguez, A., & Pontious, D. (2022). Quitting thresholds in visual search are impacted by target present detection times but not their variability. *Attention, Perception, & Psychophysics*, *84*(8), 2461–2471. https://doi.org/10.3758/s13414-022-02591-3

Benjamini, & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Benson, N. C., Kupers, E. R., Barbot, A., Carrasco, M., & Winawer, J. (2021). Cortical magnification in human visual cortex parallels task performance around the visual field. *eLife*, *10*, e67685. https://doi.org/10.7554/eLife.67685

Berbaum, K. S., Franken, E. A., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., Behlke, F. M., Sato, Y., Lu, C. H., & el-Khoury, G. Y. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology*, *25*(2), 133–140. https://doi.org/10.1097/00004424-199002000-00006

Bercovich, E., & Javitt, M. C. (2018). Medical Imaging: From Roentgen to the Digital Revolution, and Beyond. *Rambam Maimonides Medical Journal*, *9*(4), e0034. https://doi.org/10.5041/RMMJ.10355

Blakemore, C., & Campbell, F. W. (1969). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, *203*(1), 237-260.1.

Bochud, F. O., Abbey, C. K., & Eckstein, M. P. (1999). Statistical texture synthesis of mammographic images with super-blob lumpy backgrounds. *Optics Express*, *4*(1), 33–42.

Bochud, F. O., Abbey, C. K., & Eckstein, M. P. (2004). Search for lesions in mammograms: Statistical characterization of observer responses. *Medical Physics*, *31*(1), 24–36.

Bouma, H. (1970). Interaction Effects in Parafoveal Letter Recognition. *Nature*, *226*(5241), Article 5241. https://doi.org/10.1038/226177a0

Brennan, P. C., Ganesan, A., Eckstein, M. P., Ekpo, E. U., Tapia, K., Mello-Thoms, C., Lewis, S., & Juni, M. Z. (2019). Benefits of Independent Double Reading in Digital Mammography: A Theoretical Evaluation of All Possible Pairing Methodologies. *Academic Radiology*, *26*(6), 717–723. https://doi.org/10.1016/j.acra.2018.06.017

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *12*(12), 4745–4765. https://doi.org/10.1523/JNEUROSCI.12-12-04745.1992

Burgess, A. E., Jacobson, F. L., & Judy, P. F. (2001). Human observer detection experiments with mammograms and power-law noise. *Medical Physics*, *28*(4), 419–437. https://doi.org/10.1118/1.1355308

Burgess, A. E., Li, X., & Abbey, C. K. (1997). Visual signal detectability with two noise components: Anomalous masking effects. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *14*(9), 2420–2442.

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601

Calhoun, P. S., Kuszyk, B. S., Heath, D. G., Carley, J. C., & Fishman, E. K. (1999). Three-dimensional volume rendering of spiral CT data: Theory and method. *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, *19*(3), 745–764. https://doi.org/10.1148/radiographics.19.3.g99ma14745

Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, *42*(8), 949–967.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1980). An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, *9*(3), 339–344.

Carrasco, M., Talgar, C. P., & Cameron, E. L. (2001). Characterizing visual performance fields: Effects of transient covert attention, spatial frequency, eccentricity, task and set size. *Spatial Vision*, *15*(1), 61–75.

Castella, C., Kinkel, K., Descombes, F., Eckstein, M. P., Sottas, P.-E., Verdun, F. R., & Bochud, F. O. (2008). Mammographic texture synthesis: Second-generation clustered lumpy backgrounds using a genetic algorithm. *Optics Express*, *16*(11), 7595–7607.

Chen, H., De, P., Hu, Y. (Jeffrey), & Hwang, B.-H. (2014). Wisdom of Crowds: The Value of Stock Opinions Transmitted Through Social Media. *The Review of Financial Studies*, *27*(5), 1367–1403. https://doi.org/10.1093/rfs/hhu001

Chen, L., Abbey, C. K., Nosratieh, A., Lindfors, K. K., & Boone, J. M. (2012). Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies. *Medical Physics*, *39*(3), 1435–1441. https://doi.org/10.1118/1.3685462

Chong, A., Weinstein, S. P., McDonald, E. S., & Conant, E. F. (2019). Digital Breast Tomosynthesis: Concepts and Clinical Practice. *Radiology*, *292*(1), 1–14. https://doi.org/10.1148/radiol.2019180760

Chow, L. S., & Paramesran, R. (2016). Review of medical image quality assessment. *Biomedical Signal Processing and Control*, *27*, 145–154. https://doi.org/10.1016/j.bspc.2016.02.006

Chun, M. M., & Wolfe, J. M. (1996). Just Say No: How Are Visual Searches Terminated When There Is No Target Present? *Cognitive Psychology*, *30*(1), 39–78. https://doi.org/10.1006/cogp.1996.0002

Ciatto, S., Ambrogetti, D., Bonardi, R., Catarzi, S., Risso, G., Rosselli Del Turco, M., & Mantellini, P. (2005). Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme. *Journal of Medical Screening*, *12*(2), 103–106. https://doi.org/10.1258/0969141053908285

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *CoRR*, *abs/1606.06650*. http://arxiv.org/abs/1606.06650

Conant, E. F., Toledano, A. Y., Periaswamy, S., Fotin, S. V., Go, J., Boatsman, J. E., & Hoffmeister, J. W. (2019). Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis. *Radiology. Artificial Intelligence*, *1*(4), e180096. https://doi.org/10.1148/ryai.2019180096

Curcio, C. A., & Allen, K. A. (1990). Topography of ganglion cells in human retina. *Journal of Comparative Neurology*, *300*(1), 5–25. https://doi.org/10.1002/cne.903000103

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, *292*(4), 497–523. https://doi.org/10.1002/cne.902920402

Deza, A., Surana, A., & Eckstein, M. P. (2019). Assessment of Faster R-CNN in Man-Machine Collaborative Search. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3180–3189. https://doi.org/10.1109/CVPR.2019.00330

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, *31*(4), 198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002

Dorfman, D. D., & Berbaum, K. S. (2000). A contaminated binormal model for ROC data: Part II. A formal model. *Academic Radiology*, *7*(6), 427–437. https://doi.org/10.1016/s1076-6332(00)80383-9

Drew, T., Cunningham, C., & Wolfe, J. M. (2012). When and Why Might a Computer-aided Detection (CAD) System Interfere with Visual Search? An Eye-tracking Study. *Academic Radiology*, *19*(10), 1260–1267. https://doi.org/10.1016/j.acra.2012.05.013

Drew, T., Evans, K., Võ, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in Radiology: What Can You See in a Single Glance and How Might This Guide Visual Search in Medical Images? *RadioGraphics*. https://doi.org/10.1148/rg.331125023

Drew, T., Vo, M. L. H., & Wolfe, J. M. (2013). "The invisible gorilla strikes again: Sustained inattentional blindness in expert observers." *Psychological Science*, *24*(9), 1848–1853. https://doi.org/10.1177/0956797613479386

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*(10), 3–3.

Droll, J. A., Abbey, C. K., & Eckstein, M. P. (2009). Learning cue validity through performance feedback. *Journal of Vision*, *9*(2), 18.1-23. https://doi.org/10.1167/9.2.18

Duijm, L. E. M., Groenewoud, J. H., Hendriks, J. H. C. L., & de Koning, H. J. (2004). Independent Double Reading of Screening Mammograms in the Netherlands: Effect of Arbitration Following Reader Disagreements. *Radiology*, *231*(2), 564–570. https://doi.org/10.1148/radiol.2312030665

Duncan, R. O., & Boynton, G. M. (2003). Cortical Magnification within Human Primary Visual Cortex Correlates with Acuity Thresholds. *Neuron*, *38*(4), 659–671. https://doi.org/10.1016/S0896-6273(03)00265-4

Ebner, L., Tall, M., Choudhury, K. R., Ly, D. L., Roos, J. E., Napel, S., & Rubin, G. D. (2017). Variations in the functional visual field for detection of lung nodules on chest computed tomography: Impact of nodule size, distance, and local lung complexity. *Medical Physics*, *44*(7), 3483–3490. https://doi.org/10.1002/mp.12277

Eckstein, M., Bartroff, J., Abbey, C. K., Whiting, J., & Bochud, F. (2003). Automated computer evaluation and optimization of image compression of x-ray coronary angiograms for signal known exactly detection tasks. *Optics Express*, *11*(5), 460–475. https://doi.org/10.1364/OE.11.000460

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, *11*(5). https://doi.org/10.1167/11.5.14

Eckstein, M. P., Das, K., Pham, B. T., Peterson, M. F., Abbey, C. K., Sy, J. L., & Giesbrecht, B. (2012). Neural decoding of collective wisdom with multi-brain computing. *NeuroImage*, *59*(1), 94–108. https://doi.org/10.1016/j.neuroimage.2011.07.009

Eckstein, M. P., Lago, M. A., & Abbey, C. K. (2017). The role of extra-foveal processing in 3D imaging. *Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*, *10136*, 101360E. https://doi.org/10.1117/12.2255879

Eckstein, M. P., Lago, M. A., & Abbey, C. K. (2018). Evaluation of Search Strategies for Microcalcifications and Masses in 3D Images. *Proceedings of SPIE--the International Society for Optical Engineering*, *10577*. https://doi.org/10.1117/12.2293871

Eckstein, M. P., & Whiting, J. S. (1995). Lesion detection in structured noise. *Academic Radiology*, *2*(3), 249–253.

Eckstein, M. P., Whiting, J. S., & Thomas, J. P. (1996). Detection and contrast discrimination of moving signals in uncorrelated Gaussian noise. *Medical Imaging 1996: Image Perception*, *2712*, 9–25. https://doi.org/10.1117/12.236855

Eckstein, M., Whiting, J. S., Thomas, J. P., & Shimozaki, S. (1992). A novel temporal integration of intensity. *Optical Society of America Annual Meeting (1992), Paper MS1*, MS1. https://opg.optica.org/abstract.cfm?uri=OAM-1992-MS1

Essen, D. V., Newsome, W. T., & Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research*, *24*(5), 429–448. https://doi.org/10.1016/0042-6989(84)90041-5

Fan, M., Li, Y., Zheng, S., Peng, W., Tang, W., & Li, L. (2019). Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network. *Methods*, *166*, 103–111. https://doi.org/10.1016/j.ymeth.2019.02.010

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science*, *18*(11), 943–947. https://doi.org/10.1111/j.1467-9280.2007.02006.x

Fleck, M. S., Samei, E., & Mitroff, S. R. (2010). Generalized "satisfaction of search": Adverse influences on dual-target search accuracy. *Journal of Experimental Psychology: Applied*, *16*, 60–71. https://doi.org/10.1037/a0018629

Foulsham, T., & Kingstone, A. (2013). Where have eye been? Observers can recognise their own fixations. *Perception*, *42*(10), 1085–1089. https://doi.org/10.1068/p7562

Gallas, B. D., & Brown, D. G. (2008). Reader studies for validation of CAD systems. *Neural Networks: The Official Journal of the International Neural Network Society*, *21*(2–3), 387–397. https://doi.org/10.1016/j.neunet.2007.12.013

Galton, F. (1907). Vox Populi. *Nature*, *75*(1949), 450–451. https://doi.org/10.1038/075450a0

Geijer, H., & Geijer, M. (2018). Added value of double reading in diagnostic radiology,a systematic review. *Insights into Imaging*, *9*(3), 287–301. https://doi.org/10.1007/s13244-018-0599-0

Georgian-Smith, D., Obuchowski, N. A., Lo, J. Y., Brem, R. F., Baker, J. A., Fisher, P. R., Rim, A., Zhao, W., Fajardo, L. L., & Mertelmeier, T. (2019). Can Digital Breast Tomosynthesis Replace Full-Field Digital Mammography? A Multireader, Multicase

Study of Wide-Angle Tomosynthesis. *AJR. American Journal of Roentgenology*, 1–7. https://doi.org/10.2214/AJR.18.20294

Geras, K. J., Mann, R. M., & Moy, L. (2019). Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. *Radiology*, *293*(2), 246–259. https://doi.org/10.1148/radiol.2019182627

Getty, D. J., & Green, P. J. (2007). Clinical applications for stereoscopic 3-D displays. *Journal of the Society for Information Display*, *15*(6), 377–384. https://doi.org/10.1889/1.2749323

Gifford, H. C., King, M. A., Pretorius, P. H., & Wells, R. G. (2005). A comparison of human and model observers in multislice LROC studies. *IEEE Transactions on Medical Imaging*, *24*(2), 160–169. https://doi.org/10.1109/TMI.2004.839362

Gifford, H. C., Liang, Z., & Das, M. (2016). Visual-search observers for assessing tomographic x-ray image quality. *Medical Physics*, *43*(3), 1563–1575. https://doi.org/10.1118/1.4942485

Giger, M. L., Chan, H.-P., & Boone, J. (2008). Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Medical Physics*, *35*(12), 5799–5820. https://doi.org/10.1118/1.3013555

Good, W. F., Abrams, G. S., Catullo, V. J., Chough, D. M., Ganott, M. A., Hakim, C. M., & Gur, D. (2008). Digital Breast Tomosynthesis: A Pilot Observer Study. *American Journal of Roentgenology*, *190*(4), 865–869. https://doi.org/10.2214/AJR.07.2841

Gothwal, R., Tiwari, S., & Shivani, S. (2022). Computational Medical Image Reconstruction Techniques: A Comprehensive Review. *Archives of Computational Methods in Engineering*, *29*(7), 5635–5662. https://doi.org/10.1007/s11831-022-09785-w

Gould, M. K. (2014). Lung-Cancer Screening with Low-Dose Computed Tomography. *New England Journal of Medicine*, *371*(19), 1813–1820. https://doi.org/10.1056/NEJMcp1404071

Green, D. M. (1966). *Signal detection theory and Psychophysics*. Wiley.

Green, D. M., & Swets, J. A. (1989). *Signal Detection Theory and Psychophysics*. Peninsula Pub.

Gur, D., Abrams, G. S., Chough, D. M., Ganott, M. A., Hakim, C. M., Perrin, R. L., Rathfon, G. Y., Sumkin, J. H., Zuley, M. L., & Bandos, A. I. (2009). Digital Breast Tomosynthesis: Observer Performance Study. *American Journal of Roentgenology*, *193*(2), 586–591. https://doi.org/10.2214/AJR.08.2031

Hasan, E., Eichbaum, Q., Seegmiller, A. C., Stratton, C., & Trueblood, J. S. (2023). Harnessing the wisdom of the confident crowd in medical image decision-making. *Decision*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/dec0000210

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*, 494–508.

Health, C. for D. and R. (2023). MQSA National Statistics. *FDA*. https://www.fda.gov/radiation-emitting-products/mqsa-insights/mqsa-national-statistics

Helvie, M. A. (2010). Digital Mammography Imaging: Breast Tomosynthesis and Advanced Applications. *Radiologic Clinics of North America*, *48*(5), 917–929. https://doi.org/10.1016/j.rcl.2010.06.009

Hoppe, D., & Rothkopf, C. A. (2019). Multi-step planning of eye movements in visual search. *Scientific Reports*, *9*(1), 144. https://doi.org/10.1038/s41598-018-37536-0

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*(6693), 575–577. https://doi.org/10.1038/29068

Horvat, J. V., Keating, D. M., Rodrigues-Duarte, H., Morris, E. A., & Mango, V. L. (2019). Calcifications at Digital Breast Tomosynthesis: Imaging Features and Biopsy Techniques. *Radiographics*, *39*(2), 307–318. https://doi.org/10.1148/rg.2019180124

Huda, W., Ravenel, J. G., & Scalzetti, E. M. (2002). How do radiographic techniques affect image quality and patient doses in CT? *Seminars in Ultrasound, CT and MRI*, *23*(5), 411–422. https://doi.org/10.1016/S0887-2171(02)90012-0

Hueffer, K., Fonseca, M. A., Leiserowitz, A., & Taylor, K. M. (2013). The wisdom of crowds: Predicting a weather and climate-relatedevent. *Judgment and Decision Making*, *8*(2), 91–105. https://doi.org/10.1017/S1930297500005039

Hulleman, J., & Olivers, C. N. (2017). The impending demise of the item in visual search. *Behavioral and Brain Sciences*, *40*.

Isensee, F., Petersen, J., Kohl, S. A., Jäger, P. F., & Maier-Hein, K. H. (2019). nnu-net: Breaking the spell on successful medical image segmentation. *arXiv Preprint arXiv:1904.08128*.

Ishibashi, K., Kita, S., & Wolfe, J. M. (2012). The effects of local prevalence and explicit expectations on search termination times. *Attention, Perception, & Psychophysics*, *74*(1), 115–123. https://doi.org/10.3758/s13414-011-0225-4

Juni, M. Z., & Eckstein, M. P. (2015). Flexible human collective wisdom. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1588–1611. https://doi.org/10.1037/xhp0000101

Juni, M. Z., & Eckstein, M. P. (2017). The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, *114*(21), E4306–E4315. https://doi.org/10.1073/pnas.1610732114

Kattan, M. W., O'Rourke, C., Yu, C., & Chagin, K. (2016). The Wisdom of Crowds of Doctors: Their Average Predictions Outperform Their Individual Ones. *Medical Decision Making*, *36*(4), 536–540. https://doi.org/10.1177/0272989X15581615

Kim, H., Hong, J., Lee, T., Choi, Y.-W., Kim, H. H., Chae, E. Y., Choi, W. J., & Cho, S. (2020). A synthesizing method for signal-enhanced and artifact-reduced mammogram from digital breast tomosynthesis. *Physics in Medicine and Biology*, *65*(21), 215026. https://doi.org/10.1088/1361-6560/abb31e

Kim, S. T., Kim, D. H., & Ro, Y. M. (2014). Generation of conspicuity-improved synthetic image from digital breast tomosynthesis. *2014 19th International Conference on Digital Signal Processing*, 395–399. https://doi.org/10.1109/ICDSP.2014.6900693

Klein, D., & Eckstein, M. P. (2023). Sufficient eye movement coverage of the 2D image plane might mediate under-exploration in 3D search. *Journal of Vision*, *23*(9), 5831. https://doi.org/10.1167/jov.23.9.5831

Klein, D. S., Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2023). A 2D synthesized image improves the 3D search for foveated visual systems. *IEEE Transactions on Medical Imaging*, 1–1. https://doi.org/10.1109/TMI.2023.3246005

Klein, D. S., Lago, M. A., & Eckstein, M. P. (2021). The perceptual influence of 2D synthesized images on 3D search. *Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment*, *11599*, 145–155. https://doi.org/10.1117/12.2582262

Kleiner, M., Brainard, D. H., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*, 1–16. https://doi.org/10.1068/v070821

Knotts, J., Odegaard, B., Lau, H., & Rosenthal, D. (2019). Subjective inflation: Phenomenology's get-rich-quick scheme. *Current Opinion in Psychology*, *29*, 49–55. https://doi.org/10.1016/j.copsyc.2018.11.006

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., & Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, *35*, 303–312. https://doi.org/10.1016/j.media.2016.07.007

Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic Radiology*, *3*(2), 137–144.

Krupinski, E. A. (2000). The importance of perception research in medical imaging. *Radiation Medicine*, *18*(6), 329–334.

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

Krupinski, E. A. (2011). The Role of Perception in Imaging: Past and Future. *Seminars in Nuclear Medicine*, *41*(6), 392–400. https://doi.org/10.1053/j.semnuclmed.2011.05.002

Krupinski, E. A., Berbaum, K. S., Caldwell, R. T., Schartz, K. M., & Kim, J. (2010). Long Radiology Workdays Reduce Detection and Accommodation Accuracy. *Journal of the American College of Radiology*, *7*(9), 698–704. https://doi.org/10.1016/j.jacr.2010.03.004

Kundel, H. L. (1975). Peripheral vision, structured noise and film reader error. *Radiology*, *114*(2), 269–273.

Kundel, H. L. (1989). Perception errors in chest radiography. *Seminars in Respiratory Medicine*, *10*, 203–210.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting Chest Radiographs without Visual Search. *Radiology*, *116*(3), 527–532. https://doi.org/10.1148/116.3.527

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*, *13*(3), 175–181.

Kundel, H. L., Nodine, C. F., & Krupinski, E. A. (1989). Searching for lung nodules. Visual dwell indicates locations of false-positive and false-negative decisions. *Investigative Radiology*, *24*(6), 472–478.

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782. https://doi.org/10.1073/pnas.1601827113

La Vecchia, L. (2013). The History of Research on Coronary Angiography and Coronary Angioplasty. In M. Picichè (Ed.), *Dawn and Evolution of Cardiac Procedures: Research Avenues in Cardiac Surgery and Interventional Cardiology* (pp. 145–161). Springer Milan. https://doi.org/10.1007/978-88-470-2400-7_15

Lago, M. A., Abbey, C. K., Barufaldi, B., Bakic, P. R., Weinstein, S. P., Maidment, A. D., & Eckstein, M. P. (2018). Interactions of lesion detectability and size across single-slice DBT and 3D DBT. *Medical Imaging 2018: Image Perception, Observer Performance, and Technology Assessment*, *10577*, 105770X. https://doi.org/10.1117/12.2293873

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2017). Foveated model observers to predict human performance in 3D images. *Proc.SPIE*, *10136*. https://doi.org/10.1117/12.2252952

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2019). A foveated channelized Hotelling search model predicts dissociations in human performance in 2D and 3D images. *Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment*, *10952*, 109520D. https://doi.org/10.1117/12.2511777

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2020). Foveated Model Observers for Visual Search in 3D Medical Images. *IEEE Transactions on Medical Imaging*, *PP*. https://doi.org/10.1109/TMI.2020.3044530

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2021a). Foveated Model Observers for Visual Search in 3D Medical Images. *IEEE Transactions on Medical Imaging*, *40*(3), 1021–1031. https://doi.org/10.1109/TMI.2020.3044530

Lago, M. A., Abbey, C. K., & Eckstein, M. P. (2021b). Medical image quality metrics for foveated model observers. *Journal of Medical Imaging (Bellingham, Wash.)*, *8*(4), 041209. https://doi.org/10.1117/1.JMI.8.4.041209

Lago, M. A., Jonnalagadda, A., Abbey, C. K., Barufaldi, B. B., Bakic, P. R., Maidment, A. D. A., Leung, W. K., Weinstein, S. P., Englander, B. S., & Eckstein, M. P. (2021). Under-exploration of Three-Dimensional Images Leads to Search Errors for Small Salient Targets. *Current Biology: CB*. https://doi.org/10.1016/j.cub.2020.12.029

Lago, M. A., Sechopoulos, I., Bochud, F. O., & Eckstein, M. P. (2020). Measurement of the useful field of view for single slices of different imaging modalities and targets. *Journal of Medical Imaging*, *7*(2), 022411. https://doi.org/10.1117/1.JMI.7.2.022411

Lamb, L. R., Lehman, C. D., Gastounioti, A., Conant, E. F., & Bahl, M. (2022). Artificial Intelligence (AI) for Screening Mammography, From the AJR Special Series on AI Applications. *American Journal of Roentgenology*, *219*(3), 369–380. https://doi.org/10.2214/AJR.21.27071

Levi, D. M., Klein, S. A., & Aitsebaomo, A. P. (1985). Vernier acuity, crowding and cortical magnification. *Vision Research*, *25*(7), 963–977. https://doi.org/10.1016/0042-6989(85)90207-X

Lu, Z., & Sakamoto, Y. (2018). Holographic display methods for volume data: Polygon-based and MIP-based methods. *Applied Optics*, *57*(1), A142–A149. https://doi.org/10.1364/AO.57.00A142

Lui, L., Pratt, J., & Lawrence, R. K. (2024). The effect of prevalence on distractor speeded search termination. *Psychonomic Bulletin & Review*, *31*(1), 303–311. https://doi.org/10.3758/s13423-023-02337-8

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide, 2nd ed* (pp. xix, 492). Lawrence Erlbaum Associates Publishers.

Maupu, D., Van Horn, M. H., Weeks, S., & Bullitt, E. (2005). 3D stereo interactive medical visualization. *IEEE Computer Graphics and Applications*, *25*(5), 67–71. https://doi.org/10.1109/mcg.2005.94

Mazor, M., & Fleming, S. M. (2022). Efficient search termination without task experience. *Journal of Experimental Psychology: General*, *151*(10), 2494–2510. https://doi.org/10.1037/xge0001188

Mello-Thoms, C., Dunn, S. M., Nodine, C. F., & Kundel, H. L. (2003). The perception of breast cancers-a spatial frequency analysis of what differentiates missed from reported cancers. *IEEE Transactions on Medical Imaging*, *22*(10), 1297–1306. https://doi.org/10.1109/TMI.2003.817784

Mello-Thoms, C., Hardesty, L., Sumkin, J., Ganott, M., Hakim, C., Britton, C., Stalder, J., & Maitz, G. (2005). Effects of Lesion Conspicuity on Visual Search in Mammogram Reading1. *Academic Radiology*, *12*(7), 830–840. https://doi.org/10.1016/j.acra.2005.03.068

Merkle, E. C., & Steyvers, M. (2011). A Psychological Model for Aggregating Judgments of Magnitude. In J. Salerno, S. J. Yang, D. Nau, & S.-K. Chai (Eds.), *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 236–243). Springer. https://doi.org/10.1007/978-3-642-19656-0_34

Metz, C. E. (1986). ROC Methodology in Radiologic Imaging. *Investigative Radiology*, *21*(9), 720.

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*(1), 284–289. https://doi.org/10.1177/0956797613504221

Myers, K. J., & Barrett, H. H. (1987). Addition of a channel mechanism to the ideal-observer model. *JOSA A*, *4*(12), 2447–2457.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387.

Neider, M. B., & Zelinsky, G. J. (2008). Exploring set size effects in scenes: Identifying the objects of search. *Visual Cognition*, *16*(1), 1–10. https://doi.org/10.1080/13506280701381691

Nelson, J. S., Wells, J. R., Baker, J. A., & Samei, E. (2016). How does c-view image quality compare with conventional 2D FFDM? *Medical Physics*, *43*(5), 2538. https://doi.org/10.1118/1.4947293

Nodine, C. F., & Mello-Thoms, C. (2000). *The Nature of Expertise in Radiology*. *PM79*, 859–895. https://doi.org/10.1117/3.832716.ch19

Nodine, C. F., & Mello-Thoms, C. (2010). The role of expertise in radiologic image interpretation. In *The handbook of medical image perception and techniques* (pp. 139–156).

Obuchowski, N. A., & Bullen, J. (2022). Multireader Diagnostic Accuracy Imaging Studies: Fundamentals of Design and Analysis. *Radiology*, *303*(1), 26–34. https://doi.org/10.1148/radiol.211593

Odegaard, B., Chang, M. Y., Lau, H., & Cheung, S.-H. (2018). Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *373*(1755), 20170345. https://doi.org/10.1098/rstb.2017.0345

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences*, *87*(24), 9868–9872. https://doi.org/10.1073/pnas.87.24.9868

Parker, M. G., Muhl-Richardson, A., & Davis, G. (2022). Enhanced threat detection in three dimensions: An image-matched comparison of computed tomography and dual-view X-ray baggage screening. *Applied Ergonomics*, *105*, 103834. https://doi.org/10.1016/j.apergo.2022.103834

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, *4*(12), 12. https://doi.org/10.1167/4.12.12

Peterson, M. S., Kramer, A. F., Wang, R. F., Irwin, D. E., & McCarley, J. S. (2001). Visual search has memory. *Psychological Science*, *12*(4), 287–292. https://doi.org/10.1111/1467-9280.00353

Pinto, M. C., Rodriguez-Ruiz, A., Pedersen, K., Hofvind, S., Wicklein, J., Kappler, S., Mann, R. M., & Sechopoulos, I. (2021). Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with

Single-View Wide-Angle Digital Breast Tomosynthesis. *Radiology*, *300*(3), 529–536. https://doi.org/10.1148/radiol.2021204432

Pisano, E. D., Gatsonis Constantine, Hendrick Edward, Yaffe Martin, Baum Janet K., Acharyya Suddhasatta, Conant Emily F., Fajardo Laurie L., Bassett Lawrence, D'Orsi Carl, Jong Roberta, & Rebner Murray. (2005). Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. *New England Journal of Medicine*, *353*(17), 1773–1783. https://doi.org/10.1056/NEJMoa052911

Platiša, L., Goossens, B., Vansteenkiste, E., Park, S., Gallas, B. D., Badano, A., & Philips, W. (2011). Channelized Hotelling observers for the assessment of volumetric imaging data sets. *J. Opt. Soc. Am. A*, *28*(6), 1145–1163. https://doi.org/10.1364/JOSAA.28.001145

Pokrajac, D. D., Maidment, A. D. A., & Bakic, P. R. (2012). Optimized generation of high resolution breast anthropomorphic software phantoms. *Medical Physics*, *39*(4), 2290–2302. https://doi.org/10.1118/1.3697523

Predrag R. Bakic, A. D. A. M., David D. Pokrajac. (2017). Computer simulation of the breast subcutaneous and retromammary tissue for use in virtual clinical trials. *Proc.SPIE*, *10132*. https://doi.org/10.1117/12.2255099

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Purves, D., Augustine, G. J., Fitzpatrick, D., Katz, L. C., LaMantia, A.-S., McNamara, J. O., & Williams, S. M. (2001). Functional Specialization of the Rod and Cone Systems. *Neuroscience. 2nd Edition*. https://www.ncbi.nlm.nih.gov/books/NBK10850/

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Reiner, B. I., & Krupinski, E. (2012). The Insidious Problem of Fatigue in Medical Imaging Practice. *Journal of Digital Imaging*, *25*(1), 3–6. https://doi.org/10.1007/s10278-011-9436-4

Robson, J. G., & Graham, N. (1981). Probability summation and regional variation in contrast sensitivity across the visual field. *Vision Research*, *21*(3), 409–418.

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.-J., Schilling, K., Heywang-Köbrunner, S. H., Sechopoulos, I., & Mann, R. M. (2019). Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology*, *290*(2), 305–314. https://doi.org/10.1148/radiol.2018181371

Rodriguez-Ruiz, A., Lång, K., Gubern-Merida, A., Teuwen, J., Broeders, M., Gennaro, G., Clauser, P., Helbich, T. H., Chevalier, M., Mertelmeier, T., Wallis, M. G., Andersson, I., Zackrisson, S., Sechopoulos, I., & Mann, R. M. (2019). Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *European Radiology*, *29*(9), 4825–4832. https://doi.org/10.1007/s00330-019-06186-9

Roe, C. A., & Metz, C. E. (1997). Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation. *Academic Radiology*, *4*(4), 298–303. https://doi.org/10.1016/S1076-6332(97)80032-3

Rolland, J. P., & Barrett, H. H. (1992). Effect of random background inhomogeneity on observer detection performance. *Journal of the Optical Society of America. A, Optics and Image Science*, *9*(5), 649–658.

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Medical Image Computing and Computer-Assisted Intervention (MICCAI). https://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a/

Rosenholtz, R. (2016). Capabilities and Limitations of Peripheral Vision. *Annual Review of Vision Science*, *2*, 437–457. https://doi.org/10.1146/annurev-vision-082114-035733

Rovamo, J., Leinonen, L., Laurinen, P., & Virsu, V. (1984). Temporal integration and contrast sensitivity in foveal and peripheral vision. *Perception*, *13*(6), 665–674. https://doi.org/10.1068/p130665

Rubin, G. D., Beaulieu, C. F., Argiro, V., Ringl, H., Norbash, A. M., Feller, J. F., Dake, M. D., Jeffrey, R. B., & Napel, S. (1996). Perspective volume rendering of CT and MR images: Applications for endoscopic imaging. *Radiology*, *199*(2), 321–330. https://doi.org/10.1148/radiology.199.2.8668772

Rubin, G. D., Roos, J. E., Tall, M., Harrawood, B., Bag, S., Ly, D. L., Seaman, D. M., Hurwitz, L. M., Napel, S., & Roy Choudhury, K. (2015). Characterizing Search, Recognition, and Decision in the Detection of Lung Nodules on CT Scans: Elucidation with Eye Tracking. *Radiology*, *274*(1), 276–286. https://doi.org/10.1148/radiol.14132918

Saha Roy, T., Mazumder, S., & Das, K. (2021). Wisdom of crowds benefits perceptual decision making across difficulty levels. *Scientific Reports*, *11*, 538. https://doi.org/10.1038/s41598-020-80500-0

Salz, D. A., & Witkin, A. J. (2015). Imaging in Diabetic Retinopathy. *Middle East African Journal of Ophthalmology*, *22*(2), 145–150. https://doi.org/10.4103/0974-9233.151887

Samajdar, T., & Quraishi, Md. I. (2015). Analysis and Evaluation of Image Quality Metrics. In J. K. Mandal, S. C. Satapathy, M. Kumar Sanyal, P. P. Sarkar, & A. Mukhopadhyay (Eds.), *Information Systems Design and Intelligent Applications* (pp. 369–378). Springer India. https://doi.org/10.1007/978-81-322-2247-7_38

Samulski, M., Hupse, R., Boetes, C., Mus, R. D. M., den Heeten, G. J., & Karssemeijer, N. (2010). Using computer-aided detection in mammography as a decision support. *European Radiology*, *20*(10), 2323–2330. https://doi.org/10.1007/s00330-010-1821-8

Schofield, R., King, L., Tayal, U., Castellano, I., Stirrup, J., Pontana, F., Earls, J., & Nicol, E. (2020). Image reconstruction: Part 1 – understanding filtered back projection, noise and image acquisition. *Journal of Cardiovascular Computed Tomography*, *14*(3), 219–225. https://doi.org/10.1016/j.jcct.2019.04.008

Seah, J. C. Y., Tang, C. H. M., Buchlak, Q. D., Holt, X. G., Wardman, J. B., Aimoldin, A., Esmaili, N., Ahmad, H., Pham, H., Lambert, J. F., Hachey, B., Hogg, S. J. F., Johnston, B. P., Bennett, C., Oakden-Rayner, L., Brotchie, P., & Jones, C. M. (2021). Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: A retrospective, multireader multicase study. *The*

Lancet Digital Health, *3*(8), e496–e506. https://doi.org/10.1016/S2589-7500(21)00106-0

Sechopoulos, I. (2013). A review of breast tomosynthesis. Part I. The image acquisition process. *Medical Physics*, *40*(1). https://doi.org/10.1118/1.4770279

Sen, A., & Gifford, H. C. (2016). Accounting for anatomical noise in search-capable model observers for planar nuclear imaging. *Journal of Medical Imaging*, *3*(1), 015502–015502.

Sharma, B., Martin, A., Stanway, S., Johnston, S. R. D., & Constantinidou, A. (2012). Imaging in oncology—Over a century of advances. *Nature Reviews Clinical Oncology*, *9*(12), 728–737. https://doi.org/10.1038/nrclinonc.2012.195

Shi, Z., Allenmark, F., Zhu, X., Elliott, M. A., & Müller, H. J. (2020). To quit or not to quit in dynamic search. *Attention, Perception, & Psychophysics*, *82*(2), 799–817. https://doi.org/10.3758/s13414-019-01857-7

Silversmith, W. (2023, August 26). *Connected-components-3d* [Python project repository]. PyPI. https://pypi.org/project/connected-components-3d/

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3D Deep Learning on Medical Images: A Review. *Sensors (Basel, Switzerland)*, *20*(18), 5097. https://doi.org/10.3390/s20185097

Skaane, P. (2017). Breast cancer screening with digital breast tomosynthesis. *Breast Cancer (Tokyo, Japan)*, *24*(1), 32–41. https://doi.org/10.1007/s12282-016-0699-y

Skaane, P., Bandos, A. I., Eben, E. B., Jebsen, I. N., Krager, M., Haakenaasen, U., Ekseth, U., Izadi, M., Hofvind, S., & Gullien, R. (2014). Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images:

Comparison with digital breast tomosynthesis with full-field digital mammographic images. *Radiology*, *271*(3), 655–663. https://doi.org/10.1148/radiol.13131391

Smith, B., Hillis, S., & Pesce, L. (2022). *_MCMCaov: Multi-Reader Multi-Case Analysis of Variance_*. (R package version 0.2.1) [R]. https://github.com/brian-j-smith/MRMCaov

Smith, B. J., & Hillis, S. L. (2020). Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities. *Proceedings of SPIE--the International Society for Optical Engineering*, *11316*, 113160K. https://doi.org/10.1117/12.2549075

Smith-Bindman, R., Miglioretti, D. L., & Larson, E. B. (2008). Rising use of diagnostic medical imaging in a large integrated health system. *Health Affairs (Project Hope)*, *27*(6), 1491–1502. https://doi.org/10.1377/hlthaff.27.6.1491

Solovey, G., Graney, G. G., & Lau, H. (2015). A decisional account of subjective inflation of visual perception at the periphery. *Attention, Perception, & Psychophysics*, *77*(1), 258–271. https://doi.org/10.3758/s13414-014-0769-1

Sorkin, R. D., & Dai, H. (1994). Signal Detection Analysis of the Ideal Group. *Organizational Behavior and Human Decision Processes*, *60*(1), 1–13. https://doi.org/10.1006/obhd.1994.1072

Sorkin, R. D., Hays, C. J., & West, R. (2001). Signal-detection analysis of group decision making. *Psychological Review*, *108*(1), 183–203.

Sorkin, R. D., West, R., & Robinson, D. E. (1998). Group performance depends on the majority rule. *Psychological Science*, *9*(6), 456–463.

Stewart, E. E. M., Valsecchi, M., & Schütz, A. C. (2020). A review of interactions between peripheral and foveal vision. *Journal of Vision*, *20*(12), 2. https://doi.org/10.1167/jov.20.12.2

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, *11*(5), 13. https://doi.org/10.1167/11.5.13

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*(2), 181–198.

Tanner Jr, W. P., & Birdsall, T. G. (1958). Definitions of d' and η as psychophysical measures. *The Journal of the Acoustical Society of America*, *30*(10), 922–928.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), 4.1-17. https://doi.org/10.1167/7.14.4

Taylor-Phillips, S., & Stinton, C. (2019a). Double reading in breast cancer screening: Considerations for policy-making. *The British Journal of Radiology*, 20190610. https://doi.org/10.1259/bjr.20190610

Taylor-Phillips, S., & Stinton, C. (2019b). Fatigue in radiology: A fertile area for future research. *The British Journal of Radiology*, *92*(1099), 20190043. https://doi.org/10.1259/bjr.20190043

Tiley, J. (1969). The American Jury. By Harry KalvenJr. and Hans Zeisel with the Collaboration of Thomas Callahan and Philip Ennis. [Boston: Little Brown and

Company. 1967. x. and (with index) 559 pp. £5 10s. net]. *The Cambridge Law Journal*, *27*(1), 141–144. https://doi.org/10.1017/S0008197300089017

Tuddenham, W. J. (1962). Visual Search, Image Organization, and Reader Error in Roentgen Diagnosis. *Radiology*, *78*(5), 694–704. https://doi.org/10.1148/78.5.694

Tuten, W. S., & Harmening, W. M. (2021). Foveal vision. *Current Biology*, *31*(11), R701–R703. https://doi.org/10.1016/j.cub.2021.03.097

Uematsu, T., Nakashima, K., Harada, T. L., Nasu, H., & Igarashi, T. (2023). Comparisons between artificial intelligence computer-aided detection synthesized mammograms and digital mammograms when used alone and in combination with tomosynthesis images in a virtual screening setting. *Japanese Journal of Radiology*, *41*(1), 63–70. https://doi.org/10.1007/s11604-022-01327-5

Van Zandt, T., & Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics*, *53*(5), 563–580. https://doi.org/10.3758/BF03205204

Vater, C., Wolfe, B., & Rosenholtz, R. (2022). Peripheral vision in real-world tasks: A systematic review. *Psychonomic Bulletin & Review*, *29*(5), 1531–1557. https://doi.org/10.3758/s13423-022-02117-w

Võ, M. L.-H., Aizenman, A. M., & Wolfe, J. M. (2016). You think you know where you looked? You better look again. *Journal of Experimental Psychology. Human Perception and Performance*, *42*(10), 1477–1481. https://doi.org/10.1037/xhp0000264

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*. https://doi.org/10.3389/fnhum.2019.00213

Watson, A. B. (1982). Summation of grating patches indicates many types of detector at one retinal location. *Vision Research*, *22*(1), 17–25. https://doi.org/10.1016/0042-6989(82)90162-6

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*(1), 21. https://doi.org/10.1186/s41235-019-0171-6

Wolf, M., Krause, J., Carney, P. A., Bogart, A., & Kurvers, R. H. J. M. (2015). Collective Intelligence Meets Medical Decision-Making: The Collective Outperforms the Best Radiologist. *PLoS ONE*, *10*(8), e0134269. https://doi.org/10.1371/journal.pone.0134269

Wolfe, J. M. (2012). When do I quit? The search termination problem in visual search. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, *59*, 183–208.

Wolfe, J. M. (2021). Guided Search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01859-9

Wolfe, J. M., Evans, K. K., Drew, T., Aizenman, A., & Josephs, E. (2016). HOW DO RADIOLOGISTS USE THE HUMAN SEARCH ENGINE? *Radiation Protection Dosimetry*, *169*(1–4), 24–31. https://doi.org/10.1093/rpd/ncv501

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440. https://doi.org/10.1038/435439a

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General*, *136*, 623–638. https://doi.org/10.1037/0096-3445.136.4.623

Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology: CB*, *20*(2), 121–124. https://doi.org/10.1016/j.cub.2009.11.066

Wood, B. P. (1999). Visual Expertise. *Radiology*, *211*(1), 1–3. https://doi.org/10.1148/radiology.211.1.r99ap431

Wu, C.-C., & Wolfe, J. M. (2019). Eye Movements in Medical Image Perception: A Selective Review of Past, Present and Future. *Vision*, *3*(2), 32.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yang, L., Ene, I. C., Arabi Belaghi, R., Koff, D., Stein, N., & Santaguida, P. (Lina). (2022). Stakeholders' perspectives on the future of artificial intelligence in radiology: A scoping review. *European Radiology*, *32*(3), 1477–1495. https://doi.org/10.1007/s00330-021-08214-z

Yao, J., & Barrett, H. H. (1992). Predicting human performance by a channelized Hotelling observer model. *Mathematical Methods in Medical Imaging*, *1768*, 161–168. https://doi.org/10.1117/12.130899

Yu, L., Chen, B., Kofler, J. M., Favazza, C. P., Leng, S., Kupinski, M. A., & McCollough, C. H. (2017). Correlation between a 2D channelized Hotelling observer and human observers in a low-contrast detection task with multislice reading in CT. *Medical Physics*, *44*(8), 3990–3999.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–835. https://doi.org/10.1037/a0013118

Zhang, Y., Pham, B., & Eckstein, M. P. (2004a). Evaluation of JPEG 2000 encoder options: Human and model observer detection of variable signals in X-ray coronary angiograms. *IEEE Transactions on Medical Imaging*, *23*(5), 613–632.

Zhang, Y., Pham, B. T., & Eckstein, M. P. (2004b). Automated optimization of JPEG 2000 encoder options based on model observer performance for detecting variable signals in X-ray coronary angiograms. *IEEE Transactions on Medical Imaging*, *23*(4), 459–474. https://doi.org/10.1109/TMI.2004.824153

Zhang, Y., Pham, B. T., & Eckstein, M. P. (2007). Evaluation of internal noise methods for Hotelling observer models. *Medical Physics*, *34*(8), 3312–3322.

Zhou, L., Fan, M., Hansen, C., Johnson, C. R., & Weiskopf, D. (2022). A Review of Three-Dimensional Medical Image Visualization. *Health Data Science*, *2022*, 9840519. https://doi.org/10.34133/2022/9840519

Zhou, W., & Eckstein, M. P. (2022). A deep Q-learning method for optimizing visual search strategies in backgrounds of dynamic noise. *Medical Imaging 2022: Image*

Perception, Observer Performance, and Technology Assessment, *12035*, 60–67. https://doi.org/10.1117/12.2613133

Zuley, M. L., Guo, B., Catullo, V. J., Chough, D. M., Kelly, A. E., Lu, A. H., Rathfon, G. Y., Lee Spangler, M., Sumkin, J. H., Wallace, L. P., & Bandos, A. I. (2014). Comparison of two-dimensional synthesized mammograms versus original digital mammograms alone and in combination with tomosynthesis images. *Radiology*, *271*(3), 664–671. https://doi.org/10.1148/radiol.13131530

VII. Appendix

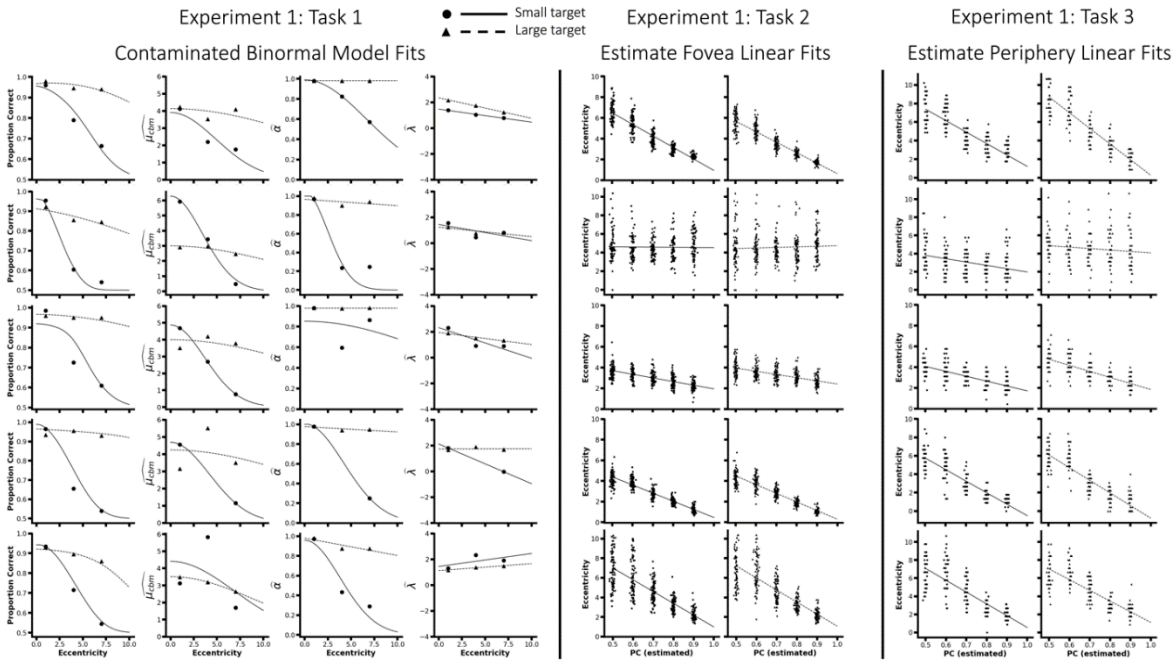

Figure A.1. Raw data and individual fits to targets for computing the three types of UFOVs for the 5 participants not shown in Figure 2.2. Each row corresponds to data from a different participant. On the left-hand side of the figure, the CBM model is applied to the proportion correct data for the two targets in task 1 of Experiment 1. A secondary fit is applied to the 3 point estimates estimated at the three tested eccentricities. In the 4th row, only two estimates for the CBM model are shown for the small target because a single rating of 1 on a target-present trial at eccentricity 4 prevented the CBM model from fitting the data. The linear fits to the raw estimates from task 2 in Experiment 1 are shown in the middle. On the right-hand side of the figure, the raw estimates and linear fits to the two targets are shown for task 3 in Experiment 1. Triangle points and dotted lines represent large target data and fits, respectively. Circular scatter points and solid lines depict small target data and fits, respectively.

Target type	Parameter	95% confidence interval	
		# of trials = 10	# of trials = 20
Small	β_o	[6.5122, 11.2751]	[6.5099, 11.2441]
	β_1	[-0.1032, -0.0394]	[-0.1032, -0.0393]
Large	β_o	[6.2943, 11.1367]	[6.3146, 11.1138]
	β_1	[-0.1026, -0.0349]	[-0.1025, -0.0350]

Table A.1. Power analysis for number of estimates needed for Experiment 1, Task 3 based on results from Experiment 1, Task 2. 95% Cis for the simple linear fit parameters for the small and large target in Experiment 1, Task 2 are shown for sampling 10 trials (with replacement) versus 20 trials (with replacement).

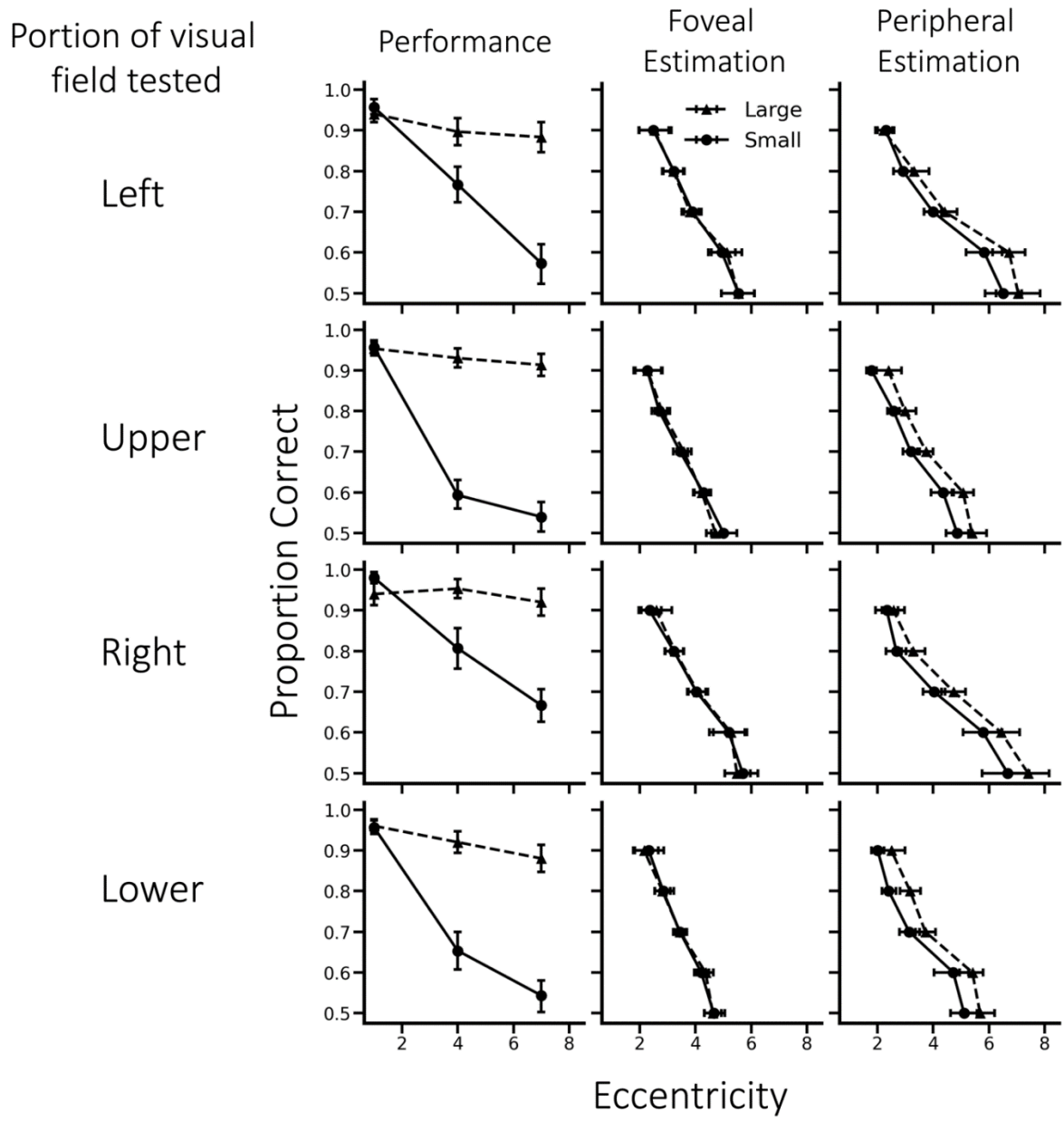

Figure A.2. Mean performance and estimation data from Tasks 1-3 in Experiment 1. From top to bottom, we move clockwise across the visual field, showing how performance and estimation change along the vertical and horizontal meridians of the visual field. Error bars represent 68% bootstrap resampling confidence intervals.

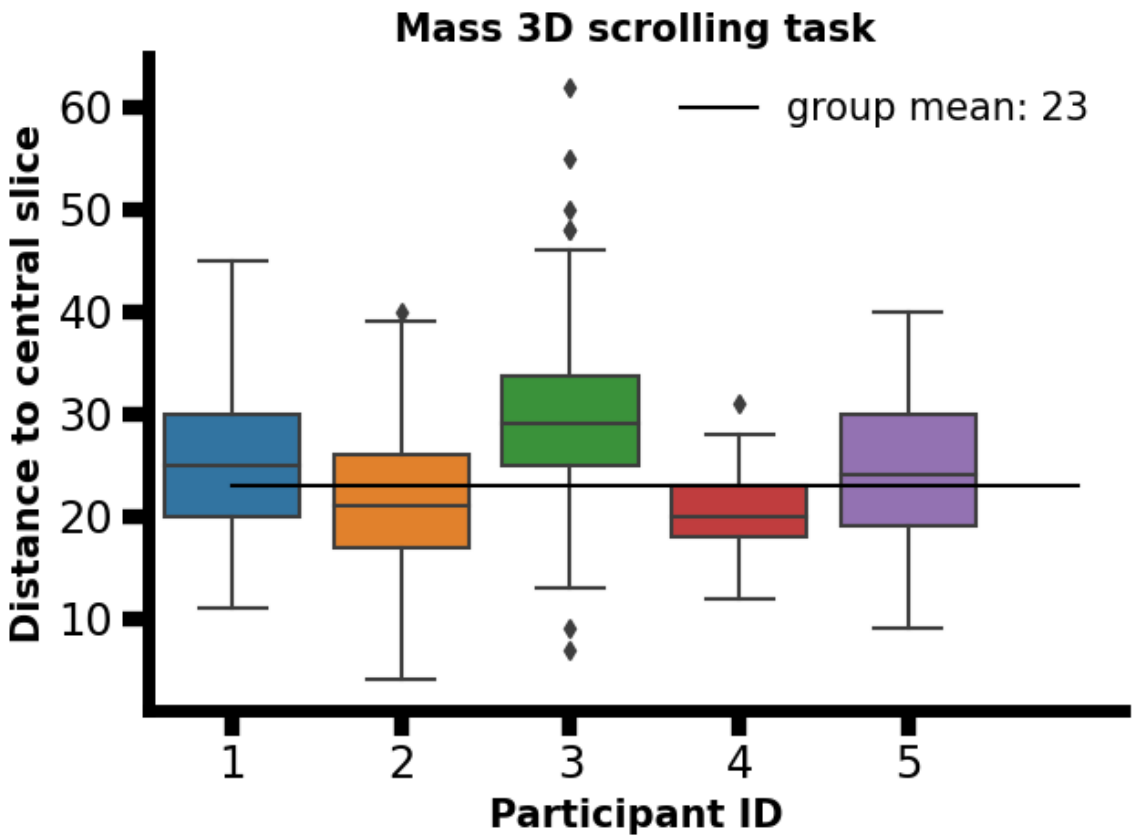

Figure A.3. Control experiment with a group of 5 new observers. Each unique colored boxplot depicts the data from 100 trials for a single observer. The dependent variable, distance to the central slice, was used to determine the threshold for mass hit rate localized and mass search errors in 3D. In particular, the mean slice distance of 23, computed across the five median values (horizontal lines within each boxplot), was used as the threshold.

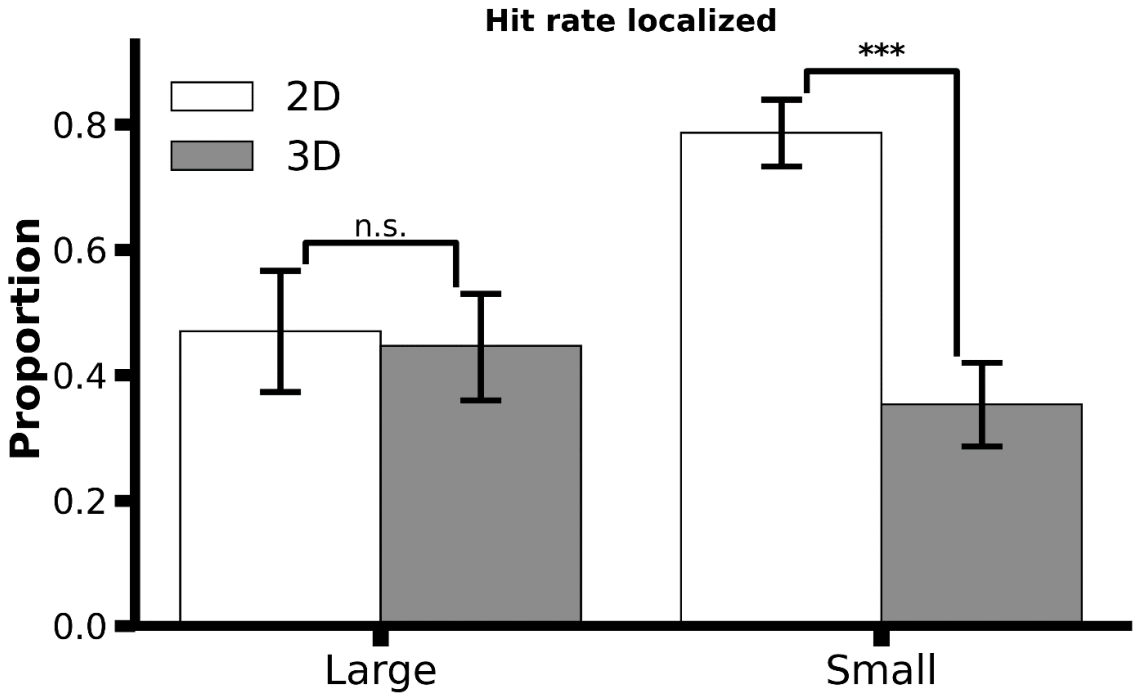

Figure A.4. The average hit rate localized is depicted for the large (left) and small (right) targets. White bars denote the hit rate localized in the 2D search, and the gray bars denote the hit rate localized in the 3D search. All error bars represent 68% bootstrap confidence intervals. "***" means $\mathbf{p < 5e^{-5}}$ and "n.s." represents non-significant results.

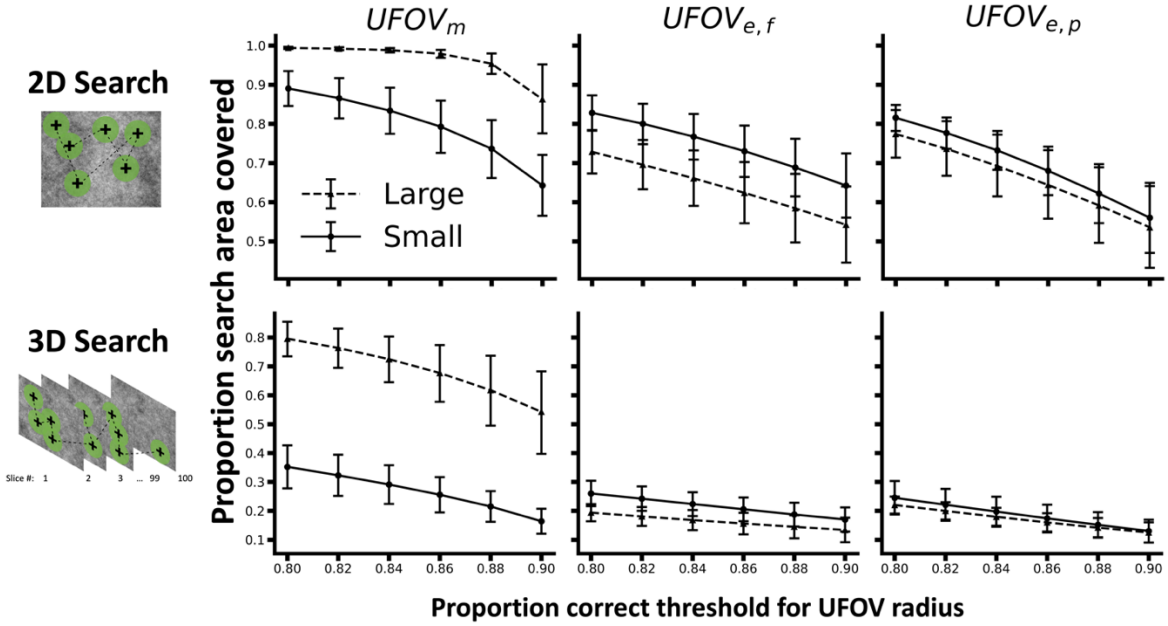

Figure A.5. Comparing the average PAC while searching for the large target (triangle points and dotted line) to the average PAC while searching for the small target for various UFOV sizes. The X-axis represents the proportion correct threshold, which determines the size of the UFOV area. The top row depicts the average PAC comparison between targets in the 2D search, and the bottom row conveys the same information as the 3D search. The three columns represent the PAC in 2D or 3D using different types of UFOVs for each target

(left-measured UFOV; middle-estimated UFOV in the fovea; right-estimated UFOV in the periphery). Error bars represent 68% bootstrap resampling confidence intervals.

Figure #	subplot	Dependent variable	Comparison group 1	Comparison group 2	Delta Mean (c1 – c2)	p-value
2.8	a, top	PAC 3D w/ $UFOV_S$	Large target	Small target	-0.05859	0.00711
		PAC 3D w/ $UFOV_{PF}$			-0.06152	0.09686
		PAC 3D w/ $UFOV_{PP}$			-0.02140	0.67260
		PAC 3D w/ $UFOV_E$			0.44072	P < 5e-5
	a, bottom	PAC 2D w/ $UFOV_S$	Large target	Small target	-0.11377	0.03783
		PAC 3D w/ $UFOV_{PF}$			-0.10472	0.13788
		PAC 3D w/ $UFOV_{PP}$			-0.04053	0.57779
		PAC 3D w/ $UFOV_E$			0.12660	P < 5e-5
	b, top	PAC ratio 3D (Large / Small)	$UFOV_E$	$UFOV_S$	2.23312	P < 5e-5
				$UFOV_{PF}$	2.19500	P < 5e-5
				$UFOV_{PP}$	1.95697	P < 5e-5
			$UFOV_{PP}$	$UFOV_{PF}$	0.23804	0.00775
				$UFOV_S$	0.27616	0.01400
			$UFOV_S$	$UFOV_{PF}$	-0.03812	0.50129
	b, bottom	PAC ratio 2D (Large / Small)	$UFOV_E$	$UFOV_S$	0.31000	P < 5e-5
				$UFOV_{PF}$	0.28819	0.00657
				$UFOV_{PP}$	0.22073	0.00680
			$UFOV_{PP}$	$UFOV_{PF}$	0.06745	0.04985
				$UFOV_S$	0.08926	0.09640
			$UFOV_S$	$UFOV_{PF}$	-0.02181	0.56222
2.9	a, top	PAC w/ $UFOV_S$ Large	2D	3D	0.50042	P < 5e-5
				2D plane of 3D	0.05564	0.03423
		PAC w/ $UFOV_{PF}$ Large		3D	0.51507	P < 5e-5
				2D plane of 3D	0.06203	0.03507
		PAC w/ $UFOV_{PP}$ Large		3D	0.53668	P < 5e-5
				2D plane of 3D	0.07048	0.03092
		PAC w/ $UFOV_E$ Large		3D	0.22837	P < 5e-5
				2D plane of 3D	0.01745	0.06563
	a, bottom	PAC w/ $UFOV_S$ Small	2D	3D	0.55560	P < 5e-5
				2D plane of 3D	-0.09026	0.04194
		PAC w/ $UFOV_{PF}$ Small		3D	0.55827	P < 5e-5
				2D plane of 3D	-0.08593	0.03408
		PAC w/ $UFOV_{PP}$ Small		3D	0.55581	P < 5e-5
				2D plane of 3D	-0.07538	0.05204

		PAC w/ $UFOV_E$ Small		3D	0.54250	P < 5e-5
				2D plane of 3D	-0.02480	0.26980
	b, top	PAC ratio w/ $UFOV_S$ Large	2D / 3D	2D / 2D plane of 3D	3.64111	P < 5e-5
		PAC ratio w/ $UFOV_{PF}$ Large			3.10566	P < 5e-5
		PAC ratio w/ $UFOV_{PP}$ Large			2.80901	P < 5e-5
		PAC ratio w/ $UFOV_E$ Large			0.34831	P < 5e-5
	b, bottom	PAC ratio w/ $UFOV_S$ Small	2D / 3D	2D / 2D plane of 3D	3.28365	P < 5e-5
		PAC ratio w/ $UFOV_{PF}$ Small			2.77643	P < 5e-5
		PAC ratio w/ $UFOV_{PP}$ Small			3.13232	P < 5e-5
		PAC ratio w/ $UFOV_E$ Small			2.46999	P < 5e-5

Table A2. Differences in the mean PAC or ratio of PACs for the various predictions made by the general hypothesis concerning the area covered by the UFOV as a stopping criterion. We report the most relevant findings in the Results section of the main text but include this table for completeness.