

# Comparative observer effects in 2D and 3D localization tasks

Craig K. Abbey,\* Miguel A. Lago<sup>✉</sup>, and Miguel P. Eckstein

University of California Santa Barbara, Department of Psychological and Brain Sciences,  
Santa Barbara, United States

## Abstract

**Purpose:** Three-dimensional “volumetric” imaging methods are now a common component of medical imaging across many imaging modalities. Relatively little is known about how human observers localize targets masked by noise and clutter as they scroll through a 3D image and how it compares to a similar task confined to a single 2D slice.

**Approach:** Gaussian random textures were used to represent noisy volumetric medical images. Subjects were able to freely inspect the images, including scrolling through 3D images as part of their search process. A total of eight experimental conditions were evaluated (2D versus 3D images, large versus small targets, power-law versus white noise). We analyze performance in these experiments using task efficiency and the classification image technique.

**Results:** In 3D tasks, median response times were roughly nine times longer than 2D, with larger relative differences for incorrect trials. The efficiency data show a dissociation in which subjects perform with higher statistical efficiency in 2D tasks for large targets and higher efficiency in 3D tasks with small targets. The classification images suggest that a critical mechanism behind this dissociation is an inability to integrate across multiple slices to form a 3D localization response. The central slices of 3D classification images are remarkably similar to the corresponding 2D classification images.

**Conclusions:** 2D and 3D tasks show similar weighting patterns between 2D images and the central slice of 3D images. There is relatively little weighting across slices in the 3D tasks, leading to lower task efficiency with respect to the ideal observer.

© The Authors. Published by SPIE under a Creative Commons Attribution 4.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JMI.8.4.041206](https://doi.org/10.1117/1.JMI.8.4.041206)]

**Keywords:** observer templates; classification images; localization tasks; volumetric imaging.

Paper 20328SSR received Dec. 2, 2020; accepted for publication Feb. 22, 2021; published online Mar. 18, 2021.

## 1 Introduction

Three-dimensional “volumetric” images are widely used in medical imaging for many purposes and across various imaging modalities. Volumetric images are appealing at a fundamental level because the 3D spatial relationships present in the body can be faithfully represented in the image up to the practical limits of contrast, resolution, and noise.<sup>1,2</sup> However, even with the development of stereo and holographic display techniques,<sup>3–5</sup> 3D images are typically displayed on a 2D monitor, which necessitates some method of accommodating this dimensionality mismatch. Many techniques for image display have been developed, ranging from surface rendering and fly-through approaches, to simultaneous multiview display.<sup>6–8</sup> Nonetheless, it is not uncommon for volumetric images to be read in a clinical setting by simply scrolling through a “stack” of 2D sections.

Scrolling replaces one of the spatial dimensions of a 3D image by mapping it into a temporal component where the reader controls the scrolling rate and direction as they search a 3D image for some target of interest. This has many potential consequences. In this work, we are interested

---

\*Address all correspondence to Craig K. Abbey, [ckabbey@ucsb.edu](mailto:ckabbey@ucsb.edu)

in what happens when the target of interest is spread across multiple sections of the 3D image in the presence of masking noise. In principle, the most effective way to find such a target will involve integrating information across these 2D sections. It is of interest at a fundamental level to know how human observers perform such an integration. At a more practical level, it is often the case that task-based psychophysical assessments of image quality in volumetric imaging modalities replace a fully 3D task with a simpler (and faster) 2D task in a single slice (e.g., Refs. 9–12). Here, the question is whether the restriction to a single “slice” image fundamentally changes the way that human subjects perform the task, potentially biasing the results of such studies. The experiments reported here are intended to make contributions to both questions.

Since our motivation is not specific to any particular (3D) imaging modality, our approach is based on generic simulated images. Simulated images have the advantage of being experimentally controllable and well characterized statistically. Both of these qualities are important for the analyses we perform. Image simulations have a long history of use establishing observer effects that impact the fields of medical image perception and vision science. Some examples of this are characterizations of visual efficiency in noise,<sup>13–15</sup> observer adaptation to image correlations,<sup>16–20</sup> internal noise,<sup>21–23</sup> and the effect of different types of tasks.<sup>24–27</sup> All of these works have used 2D simulated images to evaluate properties of human observers. There have been far fewer studies comparing and modeling observer effects between 2D and 3D images, with some notable examples<sup>28–33</sup> nonetheless, which makes the simulated-image approach more appealing for this purpose.

We investigate integration across multiple 2D sections of a volumetric image using a forced-localization task to evaluate and compare spatial weighting in noise-limited 2D and 3D images, where user-controlled scrolling is used to navigate through the slices of 3D images. The stimuli are constructed so that an ideal observer (IO) is theoretically and computationally tractable,<sup>27,34</sup> which allows us to evaluate localization efficiency as a measure of how much task-relevant information is being accessed by the human observers. The classification-image technique is used to evaluate spatial weighting used by observers to perform the tasks, which shows how information in the images is being accessed. We believe that the approach taken in this work, extending a preliminary conference report,<sup>35</sup> is a novel application of efficiency and classification images to compare 2D and 3D forced-localization tasks, which build on recent results for 2D localization tasks.<sup>27,36</sup> The noisy images we use are generated as Gaussian random fields with either a white-noise texture, as an approximation of acquisition noise, or a power-law texture, as an approximation of anatomical variability.<sup>37–40</sup> The targets to be localized are spheres (disks in 2D) of two different sizes that have been filtered and downsampled to approximate the spatial-resolution properties of modern volumetric imaging x-ray CT scanners.<sup>41–43</sup>

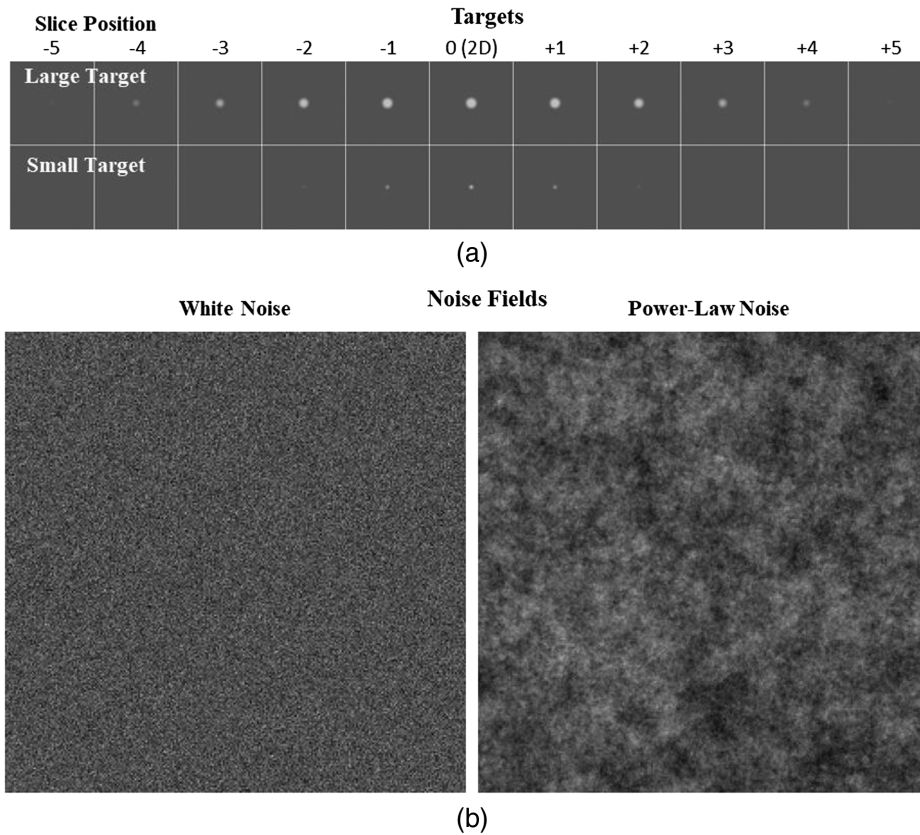
## 2 Methods

This study comprises a total of eight experimental conditions that explore localization performance across three factors: image dimension (2D and 3D), target size (large and small), and noise texture (power-law noise and white noise). Image dimension is the primary focus of the study with target size and noise texture effects giving some sense of robustness of the findings across different kinds of images.

### 2.1 Image Stimuli

All of the images used in this study are simulations generated in 3D. The 2D condition is implemented in the image display code, which only allows viewing of the slice containing the target center. The images are intended to roughly approximate a region of interest in high-resolution computed tomography (CT) imaging, with a nominal isotropic voxel size of 0.5 mm<sup>3</sup> and a total 3D image size of 256 × 256 × 256.

Figure 1(a) shows the two targets used in these experiments. Both targets are blurred spheres of constant intensity. The “large” target (Lg) has a 4 mm diameter, and the “small” target (Sm) has a 1 mm diameter. The large target extends in the *z*-direction over five slices in both directions while the small target extends over two slices in both directions. The blurring of the target



**Fig. 1** (a) Targets and noise. The 3D profile of the large target (upper row) and small target (lower row) are shown for  $\pm 5$  slices from the target center. For the 2D task, only the central slice (slice number 0) appears in the image. (b) Examples of the white noise (left) and power-law noise (right) textures used in the experiments. In the 3D tasks, these would be a single slice from a volumetric image.

profiles is intended to roughly approximate a system transfer function in an imaging context. For simplicity, we use a rotationally symmetric blurring function implemented as a filter in the FFT domain. For a radial frequency component defined as  $f = \sqrt{f_x^2 + f_y^2 + f_z^2}$ , the transfer function filter is given by a cosine roll-off function from DC ( $f = 0$ ) to Nyquist ( $f_{\text{Nyq}} = 1.0$  cyc/mm):

$$T(f) = \begin{cases} 0.5 + 0.5 \cos(\pi f / f_{\text{Nyq}}) & 0 \leq f \leq f_{\text{Nyq}} \\ 0 & f > f_{\text{Nyq}} \end{cases}. \quad (1)$$

The transfer function falls off from 1 at  $f = 0$  to 0 at  $f_{\text{Nyq}}$  with a full-width at half-max at 0.5 cyc/mm, which is roughly consistent with the transfer properties of high-resolution CT scanners.<sup>43</sup> Note that target amplitudes are defined in this work as the amplitude of the disks before filtering by the transfer function. This makes them analogous to the amplitude of lesions in tissue for the medical-imaging context.

Figure 1(b) shows sample slices for the two Gaussian noise textures used as image backgrounds. The two textures consist of white-noise (WN), in which every voxel is an independent Gaussian process, and a so-called “power-law” noise (PL) in which the power spectrum of the noise fields obeys a power-law,  $1/(f + \epsilon)^3$ , with a small offset ( $\epsilon = 0.0078$  cyc/mm) to avoid instability near  $f = 0$ . The power spectra of both processes are scaled so that the voxel standard deviation is 20 gray levels, and a mean background of 100 gray levels is used, which keeps in the images mostly well within the 8-bit display range (256 gray levels) of the monitor. Any voxels outside the 8-bit range are truncated to the nearest boundary (0 or 255).

Image backgrounds are generated by initially sampling from a standardized normal random number generator, taking the 3D FFT, multiplying by the square root of the power spectrum, inverse transforming, and then adding the mean background level. A target profile at a specified target amplitude is then added to the image background at a random location in the central  $128 \times 128 \times 128$  region of the volume, and the result is truncated to the 8-bit gray-level range of the monitor. A set of five target amplitudes are mixed across the trials. The procedure for determining these are described in the next section.

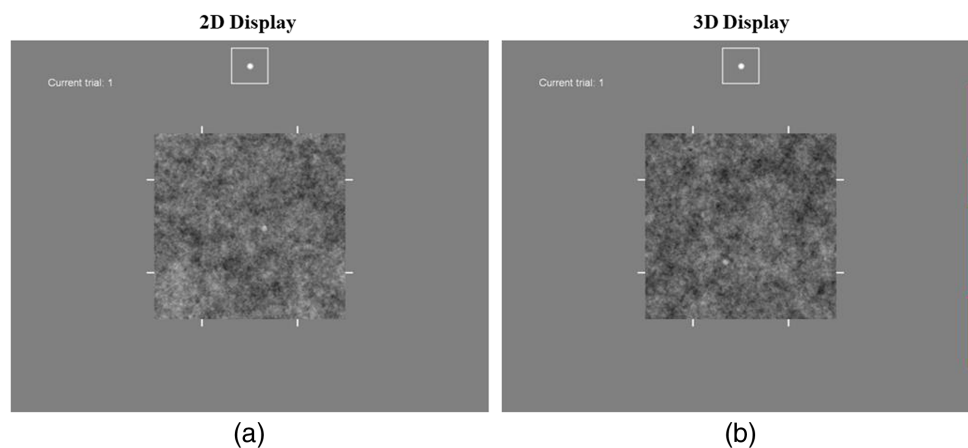
## 2.2 Forced Localization Task

Forced localization is a generalization of the multiple-alternative forced-choice paradigm. The target is always present in the image at an unknown random location, and in each trial the subject identifies the location they believe is most likely to be the target center. The response is considered correct if it falls within a distance of 6 pixels (3.6 mm radius on the display) of the actual target center.

Figure 2 shows the forced-localization interface for the 2D and 3D tasks. For the 2D tasks, a single slice is shown in the interface, as in Fig. 2(a). This slice is selected to pass through the center of the target in the  $z$  direction. The observer responds by double-clicking a mouse-driven pointer on the selected location, which must be in the central  $128 \times 128$  region of the image (i.e., inside the hash marks at the edge of the image). Responses outside of this area are ignored, and a trial lasts until a valid response is obtained.

In the 3D task shown in Fig. 2(b), the subjects need to navigate through the volume as part of the localization response. This is accomplished using a mouse click-and-drag, up or down through the  $z$  range of the 3D image. For fine tuning the slice selection, the up and down arrows on the keyboard can be used to move a single slice at a time. The scroll bar on the right side of the 3D interface is used to indicate the position of the current slice in the 3D stack. It also indicates the middle 128 slices of the  $z$  range (in green). Localization responses are only accepted within this range.

In each experimental condition, the performance is assessed in two phases. In the first “training” phase, an adaptive staircase is used to estimate the 80% correct target amplitude. We use a three-down one-up staircase in which three correct responses result in the next trial having a 15% reduced target amplitude and a single incorrect response leading to a 15% increased target amplitude. This staircase is known to oscillate around the 80% correct threshold.<sup>44</sup> The staircase starts at high amplitude to give the observer the opportunity to get familiar with the task. It typically



**Fig. 2** Localization displays. Display windows for (a) the 2D and (b) 3D stimuli are shown. The example images shown are for the large-target in power-law noise condition. A reference image of the noiseless target is displayed at the top of the window (in the box), and the trial number is shown in the upper left side of the window. The small hash marks on the edge of the image indicate the X-Y search region. On the 3D display, the scroll-bar on the right side shows the depth of the current slice in the Z direction (blue) along with the depth range (green) of the 3D search region.

takes 20 to 30 trials for the first incorrect response to be made. The staircase is run for a total of 12 reversals, in which the amplitude goes from decreasing to increasing or vice-versa. The threshold estimate is derived from the geometric mean of the target amplitude over the last eight reversals. The adaptive staircase procedure is run three times, with the final training threshold estimate being the average of the three runs.

A total of 500 forced localization trials are used for the test set, which uses five different target amplitudes that are randomly mixed throughout the trials (100 trials at each of the amplitudes). This includes the 80% correct threshold estimated from the training runs, as well as  $\pm 10\%$  of this threshold and  $\pm 20\%$  of this threshold. The range of amplitudes gives us some ability to assess the subjects' psychometric functions and also ensures that there will be a reasonable frequency of difficult cases leading to a sufficient number of incorrect responses for estimating a classification image. In each trial, the display software records the index of the stimulus, the target amplitude of the trial, the true location of the target, the localization response of the subject, and the reaction time from stimulus display to the recording of a valid mouse click. The true target location is given as  $x$ ,  $y$ , and  $z$  indices of the target center. The localization response is coded as  $x$ ,  $y$ , and  $z$  indices of the subject-selected image pixel. In the 2D task, the  $z$  index of the localization response is constrained to be the target  $z$  index. The proportion of correct responses (PC) is used as the measure of performance for a given amplitude. It is computed for each of the five amplitudes tested.

The experimental data were collected using a clinical review monitor (Barco Inc.) calibrated to the DICOM standard over a measured luminance range of 0.04 to 165.7  $\text{cd}/\text{m}^2$ . Images were magnified by a factor of 2 for a displayed pixel size of 0.6 mm, given the native (isotropic) pixel size of 0.3 mm. Subjects were encouraged to position themselves at a comfortable viewing distance, which was typically between 50 and 100 cm from the monitor face. For a subject at the center of this range, 21.8 pixels subtend a visual angle of 1 deg.

A total of five subjects conducted the studies reported here under an IRB-approved human subjects protocol at the authors' institution. The four 2D experiments were completed in roughly 30 to 45 min per condition, but the 3D experiments took considerably longer, requiring 3 to 4 h for each condition. The total time to complete the study for each subject was roughly 20 h, spread over multiple sessions at the workstation. Four of the subjects were naïve to the purpose of the research and compensated for their time, the other subject is the first author.

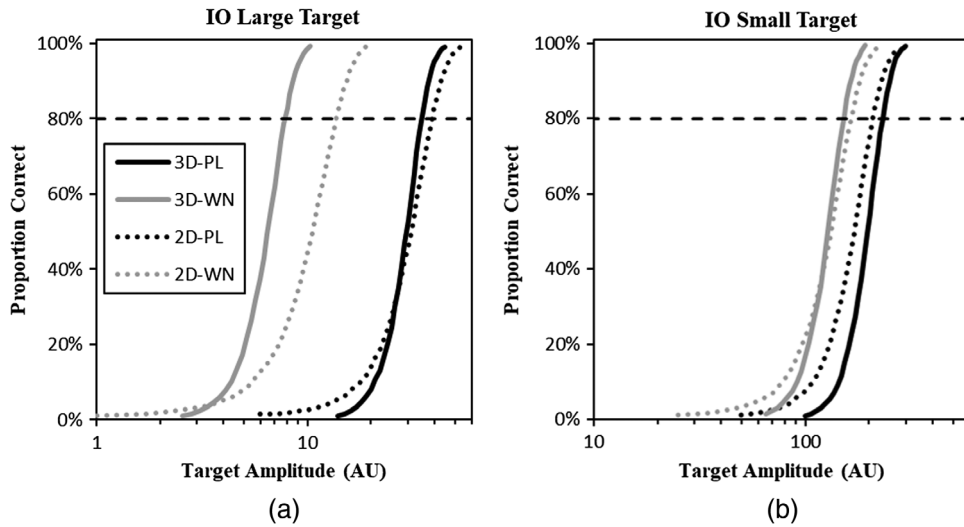
### 2.3 Ideal Observer

The Ideal Observer, described in a previous publication,<sup>27</sup> was used in the computation of efficiency. We briefly review the computations involved in evaluating the IO on a given image here. The first step involves a convolution with the prewhitened matched filter,<sup>45</sup> then exponentiation of the result (within the search region) to form a posterior distribution on target location. A second scanning operation with a 6-pixel radius disk (in 2D) or sphere (in 3D) is used to compute the posterior utility of each point in the search region. The point that maximizes this utility function over all possible locations is the IO response for the trial.

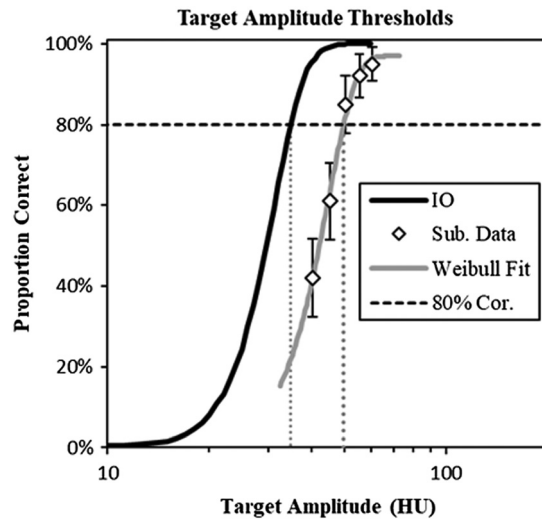
Monte-Carlo studies using many independent sample images at a given target amplitude are used to assess the performance of the IO in terms of the proportion of correct localizations (PC). Evaluations at a range of target amplitudes can be used to obtain the ideal-observer psychometric function, which shows how target amplitude affects performance in each condition. Ideal-observer psychometric functions in all eight experimental conditions are plotted in Fig. 3 using 5000 Monte-Carlo trials at each of the target amplitudes. These data are used to get ideal-observer amplitude thresholds for the efficiency computations described next.

### 2.4 Amplitude Thresholds and Efficiency

Figure 4 shows how subject data and an ideal-observer psychometric function are used to obtain an estimate of human-observer efficiency for a given experimental condition. As described above, the psychophysical experiments evaluate five different target amplitudes in each condition from which five performance levels are estimated for each subject. These points are used to fit a Weibull psychometric function,<sup>46,47</sup>  $PC(A)$ , defined as



**Fig. 3** IO psychometric functions. IO psychometric functions are shown for the tasks with (a) the large target and (b) small target. Each plot shows 3D and 2D performance in power-law noise (PL) and white noise (WN). Each point in the plot is the outcome of 5000 Monte-Carlo trials. The targeted performance level of 80% in indicated by the dashed line. Note the different ranges of the logarithmic  $x$  axis showing much lower large-target thresholds. Legend applies to both plots.



**Fig. 4** Target amplitude threshold computation. The plot shows how the 80% correct target threshold and task efficiency are computed for a set of observer performance data. The threshold amplitude ( $A_{\text{Sub}}$ ) is derived from a Weibull psychometric curve that is fit to psychometric data. An equivalent threshold for the IO ( $A_{\text{IO}}$ ) is determined from the Monte-Carlo performance evaluations. Efficiency is defined as the squared ratio of these two target amplitudes.

$$PC(A) = P_B + (1 - P_B - P_E) \left( 1 - 2^{-\left(\frac{A}{\lambda}\right)^k} \right), \quad (2)$$

where  $P_B$  is the baseline probability of a correct response (0.34% in 2D and 0.04% in 3D),  $P_E$  is the lapse rate (assumed to be 3%),  $\lambda$  is the half-rise amplitude, and  $k$  controls the steepness of the psychometric function. The  $\lambda$  and  $k$  parameters are fit using maximum likelihood, assuming observed subject PCs represent binomial proportions. Once the psychometric function has been determined, the 80% correct amplitude threshold is computed by setting  $PC(A) = 0.8$  in Eq. (2), and solving for  $A$ . This is seen in Fig. 4 as a vertical line from the intersection of the 80% correct line with the Weibull psychometric function to the  $x$  axis, defining the subject's amplitude threshold,  $A_{\text{Sub}}$ .

The 80% correct amplitude threshold for the IO is computed by a similar process from the IO psychometric data described above. Since these data are generated from many more trials than the human data (5000 trials per datum instead of 100), and a much finer sampling of amplitudes (50 instead of 5), the IO threshold is found by linear interpolation between the nearest two points, yielding  $A_{IO}$ . Efficiency with respect to the IO is then defined as the ratio<sup>48-50</sup>

$$\eta = \left( \frac{A_{IO}}{A_{Sub}} \right)^2. \quad (3)$$

## 2.5 Classification Images

Classification image analysis follows the technique described previously for forced-localization tasks.<sup>27</sup> The classification images are estimated from noise fields of the image stimuli in incorrect trials.<sup>51,52</sup> Within each condition and within each subject, these noise fields are all aligned to the (incorrect) response location and then filtered with the inverse-covariance matrix to disambiguate the effects of noise correlations. Since the images are generated from a stationary Gaussian process, this step is implemented through finite Fourier transforms and the inverse noise power spectrum. The resulting filtered noise fields are then averaged to obtain the raw classification image for each subject in each condition. For the 3D images, this process is implemented using the full 3D noise field and 3D inverse-covariance filtering. In the 2D conditions, we use the noise field of the displayed 2D slice. In this case, inverse-covariance filtering is implemented using the slice power-spectrum, which is derived from the 3D power spectrum by integrating in  $z$ . The resulting classification images are averaged across subjects for evaluating group effects of the experimental conditions.

The raw classification images can be quite noisy themselves, particularly in the power-law noise condition where low power-spectral density at higher frequencies can amplify estimation error. We use two methods to control for noise: smoothing and spatial windowing. The smoothing operation is implemented by filtering in the 2D frequency domain, with  $f_{2D} = \sqrt{f_x^2 + f_y^2}$ . For 3D classification images, smoothing is applied to each slice independently. We apply smoothing filters that are unity for  $f_{2D} < 0.5$  cyc/mm, and roll off for  $0.5 \leq f_{2D} \leq 1.0$  cyc/mm with a cosine-bell profile.

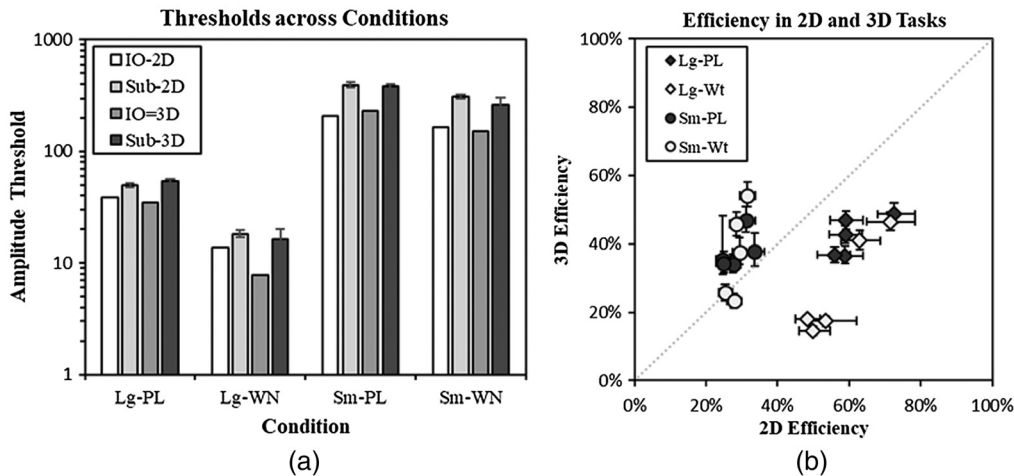
## 2.6 Scanning Models

Classification images are most readily interpreted as representing an estimate of the weights of a linear template model. This has been demonstrated analytically for detection tasks at a fixed location<sup>53-55</sup> and empirically for tasks that involve search such as the forced-localization tasks used here.<sup>27,56</sup> In localization tasks, the linear template is assumed to scan the entire search region by a convolution operation, much like the first step of the IO model described above. The localization response of the model is typically generated by taking the maximum response of the template within the search region.

When a classification image is used as the linear kernel of a scanning model, the estimation error in the classification image can bias performance of the model. Since estimation error is unlikely to be well tuned to a target profile, this bias is typically toward lower performance. To minimize this effect, we implement a number of steps to control noise in the classification images, including frequency filtering, spatial windowing, and radial averaging. These are described in Sec. 4.3.

## 3 Results

The primary analyses of the experiments are presented here, averaged over subjects. These include the observed amplitude thresholds and efficiency, response times, and classification images in each of the experimental conditions.



**Fig. 5** Amplitude thresholds and task efficiency. (a) Thresholds for each condition (Lg, large target; Sm, small target; PL, power-law noise; WN, white noise) are plotted for the IO and the average across human subjects (error bars represent a 95% confidence interval). (b) Subject efficiency is plotted as a scatterplot comparing 2D and 3D search conditions for each of the five subjects. The error bars are 95% confidence intervals generated by bootstrapping across sessions.

### 3.1 Task Performance

Figure 5 summarizes estimated amplitude thresholds for both the IO and the subjects, as well as statistical efficiency of the subjects according to Eq. (3). The amplitude thresholds in Fig. 5(a) vary considerably across the different target-size and noise-texture conditions but are relatively consistent across 2D and 3D display conditions. On average, the relative difference between subject amplitude thresholds in 2D and 3D tasks is 10.5% (min: 3.7%; max: 19.3%), and the qualitative effect of differences in target size and/or noise texture are identical. The IO thresholds are also qualitatively consistent across target size and noise texture conditions, even though the large-target white-noise condition has a 2D threshold that is 75% higher than the 3D condition.

The scatterplot of subject efficiency in Fig. 5(b) shows a clear dissociation between large targets, which are more efficiently localized in 2D, and small targets, which appear to be more efficiently localized in 3D. These differences are statistically significant (paired comparison *t*-test across subjects) in all cases except for the small-target white-noise condition. The three significant differences all survive a false-discovery rate (FDR) correction for multiple comparisons at the 5% level.<sup>57</sup> The FDR-corrected *p*-values are Lg-PL  $p < 0.013$ ; Lg-WN  $p < 0.0016$ ; Sm-PL  $p < 0.019$ ; Sm-WN  $p < 0.156$ . We will return to this dissociation in Sec. 4.

### 3.2 Response Times

Table 1 shows the response times in each condition, computed as the median response time averaged across subjects ( $\pm$  the standard deviation across subjects). Response times are given for all trials and then broken into trials in which the subjects responded correctly or trials in which the subjects responded incorrectly. Across target-size and noise-texture conditions, 3D trials take 8.9 times longer on average than 2D trials to generate a localization response. This is not surprising given the additional time needed to scroll through the search volume in 3D localization trials. Nonetheless, this larger response time difference does illustrate a substantial practical difficulty of investigating 3D image tasks.

Compared to median times for all trials, correct trials are generally somewhat faster and incorrect trials are generally substantially slower. In 2D tasks, correct trials are 5.8% faster on average and incorrect trials are 51% slower. In 3D tasks, correct trials are generally 13% faster and incorrect trials are 132% slower. It is clear that when subjects make an incorrect localization response, they have spent a relatively large amount of time searching for the target, particularly in the 3D tasks.



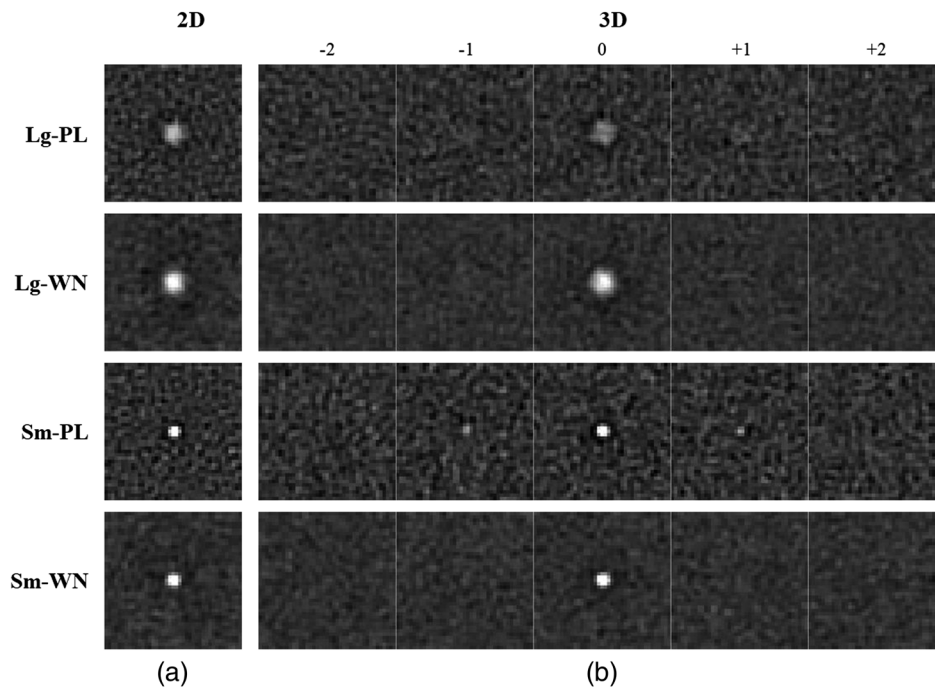
**Table 1** Median response times.

	Dim	Lg-PL	Lg-WN	Sm-PL	Sm-WN
All trials	2D	$2.1 \pm 0.6$	$1.9 \pm 0.05$	$2.2 \pm 0.6$	$1.9 \pm 0.5$
	3D	$16.5 \pm 2.5$	$17.4 \pm 5.0$	$21.5 \pm 3.7$	$17.4 \pm 2.6$
Correct trials	2D	$2.0 \pm 0.5$	$1.8 \pm 0.4$	$2.1 \pm 0.6$	$1.8 \pm 0.5$
	3D	$13.8 \pm 2.3$	$15.1 \pm 3.9$	$19.3 \pm 3.3$	$15.1 \pm 1.9$
Incorrect trials	2D	$3.6 \pm 2.0$	$2.7 \pm 1.2$	$3.4 \pm 1.8$	$2.6 \pm 1.0$
	3D	$40.3 \pm 16.7$	$38.5 \pm 21.2$	$56 \pm 21.8$	$34 \pm 8.9$

### 3.3 Classification Images

The average classification images, estimated as described in Sec. 2.5, are shown in Fig. 6. The left column of the panel in Fig. 6(a) is the 2D classification images for each target-size and noise-texture condition. The remaining portion of the panel shows the central five slices of the 3D classification images. The classification images have been frequency filtered for noise control according to the methodology described above (1 to 0.5 cyc/mm and rolled off to zero at 1 cyc/mm).

In the 2D portion of the panel, the classification images all have a center-surround profile, where a bright central region of positive weights are surrounded by a darker region of negative weights. The classification images are clearly tuned to the size of the target (i.e., larger areas of activation for larger targets). The width and magnitude of the surround appears to vary across conditions. The central slice of each 3D classification image is very similar in appearance to the 2D classification image. Off of the central slice, the activation appears to be much weaker, if it can be seen at all. There is some evidence of weak positive activation at  $\pm 1$  slice. But given that



**Fig. 6** Classification images. The average classification image across subjects is shown for each condition in (a) the 2D and the central five slices of (b) the 3D tasks (response slice  $-2$  to response slice  $+2$ ). Smoothing filters have been applied to the images.

the small signal extends over a total of five slices, and the larger target extends over 11 slices, this represents very limited use of multiple slices.

## 4 Discussion

### 4.1 Comparisons with Prior Investigations

The results of our studies can be related to findings in some earlier studies. Reiser and Nishikawa<sup>30</sup> compared 2D and 3D images in a free search task with noise structures that are very similar to what is used here (white noise and power-law noise) and targets that are closer in size to the large target in this work. They found a pronounced improvement in performance for 3D images in the white noise backgrounds, and little—if any—improvement for the power law noise. Balta et al.<sup>32</sup> also used a power-law background (with additional orientation parameters) with blurred disk targets in a signal-known-exactly task. In this case, a more realistic image formation model was used that modeled the limited angular range of digital breast tomosynthesis. They also found similar performance between 2D and 3D images, consistent with Reiser and Nishikawa.

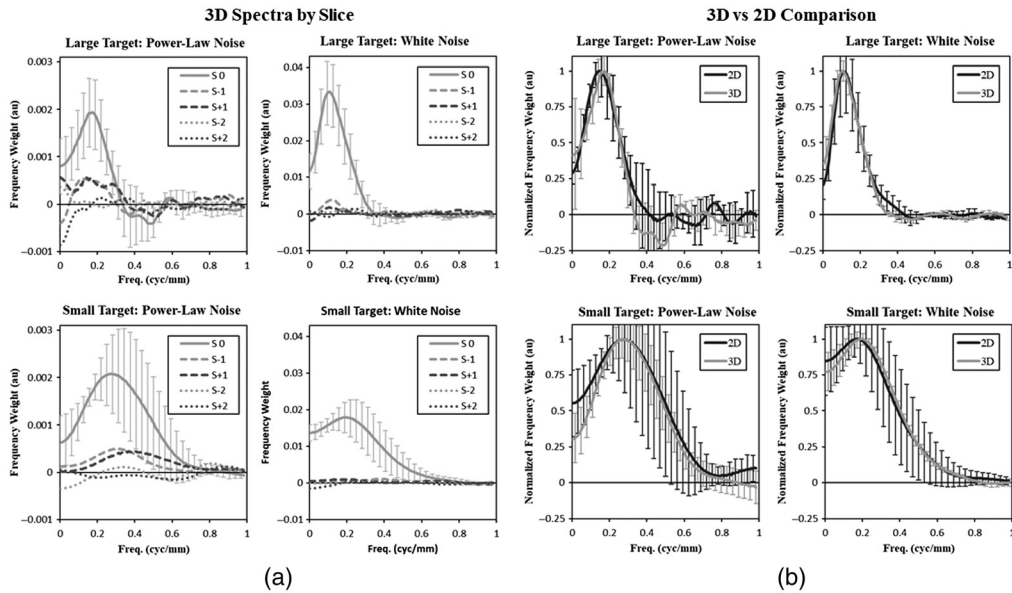
We find similar results in Fig. 5(a) for the ideal- and human-observer amplitude threshold data, although our difference is somewhat less dramatic than the finding in Reiser and Nishikawa. In white noise, the large-target amplitude thresholds drop in 3D relative to 2D, whereas in power-law noise they stay approximately the same. Thus, the absolute performance effects appear to have some robustness properties. However, Fig. 5(b) shows the importance of considering task efficiency as well. While observer performance localizing the large target is roughly equivalent in 3D and 2D images (the 3D amplitude threshold is 7% larger for power-law noise and 11% smaller for white noise), the subjects are considerably more efficient in the 2D task than the 3D task (44% more efficient in power-law noise and 108% more efficient for white noise).

### 4.2 Dissociation between Large and Small Target Efficiency

If we consider these tasks from the perspective of the threshold amplitude, shown in Fig. 5(a), then it is clear that the small targets are substantially more difficult to localize accurately than the large targets in both 2D and 3D tasks with thresholds that are 7 to 17 times larger. There are two possible reasons for this large discrepancy: (1) the tasks with small targets are inherently more difficult or (2) human observers are less effective at localizing the small targets. The efficiency values in Fig. 5(b) help disambiguate these two effects by correcting for task difficulty and therefore isolating reader performance effects. In this context, the reader results show a dissociation in which large targets are more efficiently localized in the 2D tasks and the small targets are more efficiently localized in the 3D tasks.

This finding would appear to be at odds with recent studies by Lago, Eckstein, and colleagues,<sup>33,58–60</sup> demonstrating substantial performance reductions for small targets in 3D search tasks. However, it is important to note a fundamental difference between those experiments and the results reported here. Their investigations examine the role of peripheral vision in modulating search performance in 2D and 3D images. Their images can occupy a much larger portion of the visual field than these studies (up to 30-deg visual angle). The search region used in these experiments can be mostly or entirely covered by central vision. Clinical ophthalmology texts define the fovea (including the perifovea) as occupying the central 8 deg of the visual field.<sup>61</sup> With this definition and our display procedure described, the entire  $128 \times 128$  search region fits in the fovea at a viewing distance of 76 cm or more. At a close viewing distance of 50 cm, 67% of the search region is covered by the fovea. Given the search region size and subject viewing distance, it is perhaps not surprising that we do not see evidence of peripheral vision effects.

The classification images, on the other hand, suggest that a major source of inefficiency for large targets is the lack of spatial integration across multiple slices in the 3D images, when viewed by scrolling. The spatial weights in the classification images are largely gone after the central slice. This can be seen in the off-center slices of the 3D classification images in Fig. 6.



**Fig. 7** Classification-image spectra. Spectral plots of the classification images are shown. (a) For the 3D classification images, a spectral plot from five slices is shown ( $-2$  to  $+2$  slices from response slice). Error bars representing a 95% confidence interval across subjects are plotted on the central slice (at every fourth point). Error bars on the other slices are similar in magnitude, but not shown for clarity. (b) For the comparison across 2D and 3D classification images, spectral plots from the central slice of the 3D classification image is compared to the 2D spectra after normalization so the peak frequency is 1. The 95% confidence error bars are plotted every fourth sample here as well, with an offset of two samples between the two plots for clarity.

Figure 7(a) shows the classification images in the frequency domain as the average spectral weight at each radial frequency. This gives a more quantitative comparison of the difference between the central slice and the adjacent slices. In both of these figures, there is some evidence for mild weighting of slices immediately adjacent to the central slice in the power-law noise conditions and almost no evidence for off-center weighting in the white noise condition. A failure to integrate target information across multiple slices has a greater effect on efficiency for larger targets that are spread over more slices, consistent with the efficiency results we find. This is also broadly consistent with the use of multiple views for volumetric images in the clinical context, where different views would be used to ensure 3D information is integrated into a final decision.

### 4.3 Similarity between 2D and 3D Classification Images

The 2D classification image is visually similar to the central slice of the 3D classification image, as seen in Fig. 6. Figure 7(b) shows that the average spectral weights are similar as well, with both 2D and 3D classification images adopting bandpass profiles. Table 2 quantifies these similarities in terms of the common bandpass features of peak frequency and fractional bandwidth

**Table 2** Peak frequency and fractional bandwidth for each condition.

Cond	2D	3D	2D (%)	3D (%)
Lg-PL	0.15	0.17	137	118
Lg-WN	0.11	0.10	150	177
Sm-PL	0.28	0.27	181	146
Sm-WN	0.17	0.20	223	204

(FWHM relative to peak frequency). The average relative difference between 2D and 3D conditions is  $-4\%$  for peak frequency and  $8\%$  for fractional bandwidth. For comparison, consider the average relative difference between power-law and white noise, which is  $-34\%$  for peak frequency and  $31\%$  for fractional bandwidth. Alternatively, the average relative difference between the large target and the small targets is  $75\%$  for peak frequency and  $30\%$  for fractional bandwidth. Thus, relative to other effects in these data, differences between 2D and the central slice of the 3D classification images are small.

This similarity between 2D and 3D classification images, along with the lack of substantive off-center weighting in the 3D classification images, establishes a mechanistic similarity between the 2D and 3D localization tasks. Despite the differences in image display and regardless of the search procedure used, subjects appear to be localizing targets in the 3D images as if they were looking mainly at that 2D slice. This lends some credence to the practice of evaluating 3D images using a single 2D slice, although there are many potential caveats and limitations to this statement as described below.

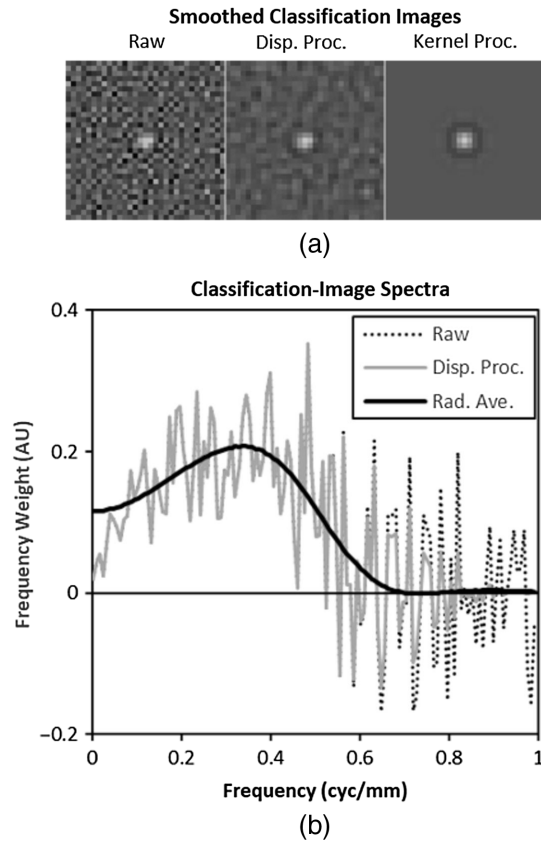
#### 4.4 Classification Images as Kernels of a Scanning Localization Model

The classification image can be interpreted as an estimate of the filter kernel<sup>27,36</sup> in the context of scanning models of localization performance. In fact, validation of classification-image estimation for localization tasks is based on generating responses from a scanning linear model and showing that the classification image accurately estimates the kernel of this model. This class of model has been used to understand search in medical images previously,<sup>62-65</sup> although the recent results of Lago et al.<sup>59,60</sup> serve as a caution when peripheral vision effects may be present. Nonetheless, the classification images can be used to understand how much of the subject's efficiency is due to the spatial weighting implemented in the scanning kernel and how much is due to other processes in the localization tasks (e.g., inefficient search or internal noise).

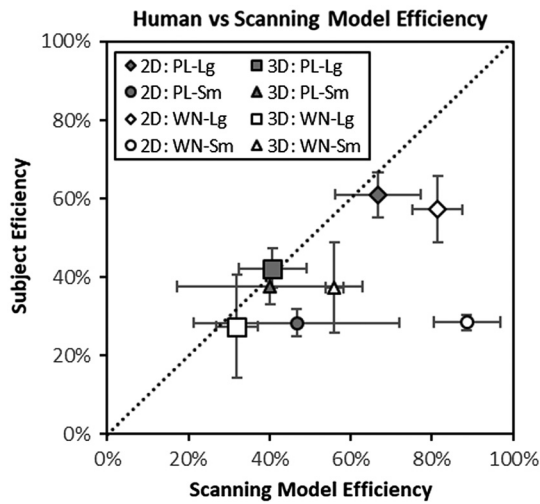
Estimation error is an important issue for implementing the classification images in scanning models. Noise in the classification image estimate will tend to reduce performance (and therefore the localization efficiency) of the model since it is unlikely that estimation error will be well tuned to a target profile. To mitigate the effects of estimation error, we use relatively aggressive filtering of the classification images based on the frequency profiles shown in Fig. 7. For the large targets, the smoothing filter extends to  $0.3$  cyc/mm before rolling off to zero with a cosine profile at  $0.6$  cyc/mm. For the small targets (which extend further into the frequency domain), the smoothing filter is constant to a frequency of  $0.6$  cyc/mm and rolls off to zero at  $1$  cyc/mm (which is identical to the filtering used in Fig. 6). In addition, radial averaging is used to smooth radial bands in the spatial domain, under the assumption of approximate rotational symmetry, and a spatial window is applied under the assumption of a relatively compact filter kernel. This spatial window is also tuned to the size of the targets. For the large targets, the spatial window is constant out to a radius of  $4$  mm and rolls off to zero at  $6$  mm with a cosine profile. For the small targets, the spatial window is constant out to a radius of  $2$  mm and rolls off to zero at  $4$  mm.

Figure 8(a) shows an example of the effects of different filtering procedures on the classification image. A raw classification image for a given subject in one of the tasks (PL-Sm) is shown along with the "display processed" version that has been frequency-filtered as in Fig. 6, and a "kernel processed" version that has been processed as described above. The kernel processed image is seen to be largely devoid of visible estimation error. For the 3D classification images, kernel processing is applied to the central three slices, with slices outside this range set to zero. Figure 8(b) shows the real component of the frequency spectrum for the various versions of the classification image. The display processed classification image is seen to have frequencies modulated starting at  $0.5$  cyc/mm and completely eliminated at  $1$  cyc/mm, consistent with the filter used to smooth the image data. The spectrum of the kernel-processed classification image is seen to have a spectrum that is generally consistent with the others, but substantially smoother.

Figure 9 shows the average subject efficiency as a function of the average efficiency of the classification-image-derived scanning models. In previous work,<sup>36</sup> task efficiency has been reasonably well modeled as kernel efficiency minus  $12.6\%$  points with a coefficient of determination ( $R^2$ ) of  $0.86$ . While that relationship seems to hold reasonably well on average in



**Fig. 8** Filtering classification images for scanning models. (a) The images and (b) frequency plots show the effect of smoothing approaches applied to the raw classification image for extracting a scanning kernel. Smoothing for display (Disp. Proc.) as in Fig. 6 is seen to remove some noise, especially at higher spatial frequencies, but not as much as “kernel processing,” which also includes radial averaging and a spatial window (see text). The spectral content of the various smoothed images is relatively similar up to no noise effects.



**Fig. 9** The average efficiency of scanning models derived from classification images is plotted against the efficiency of subjects in each of the eight tasks. Error bars represent 95% confidence intervals across subjects.

this data (average kernel efficiency minus average task efficiency is 16.5% points), the association is much weaker with  $R^2 = 0.14$ . However, one of the eight data points on the plot appears to be driving the lack of association. This point represents the 2D task with a small target and white noise (task efficiency is 28.5% and kernel efficiency is 88.6%). If we exclude this data point, association improves considerably with  $R^2 = 0.68$ .

This extreme point bears further consideration. The difference between kernel efficiency and task efficiency is more than 50% points. This suggests a relatively optimal kernel combined with substantial deficiencies in other components of task performance, such as incomplete search or internal noise. The task efficiency is relatively low compared to previous studies<sup>36</sup> that included target localization in white noise, where task efficiency was closer to 60%. It should be noted that the values reported for this condition are relatively consistent across the five subjects, ranging from 25.3% to 31.6%, so the observed value is not driven by a single outlying subject. Thus, it would seem that there may be some aspect of the stimuli or display that leads the subjects to have particularly poor performance despite an efficient kernel in this condition.

#### 4.5 Limitations

The discussion above of the extreme point in Fig. 8 indicates that there are some limitations on the interpretation of the specific conditions in this study, particularly in regards to the scanning linear kernel model. It is also important to recognize a few more general limitations in these experiments. The fact that we find little evidence of integration across multiple slices of a 3D image is likely due, at least in part, to the display procedure, which only allows the reader to view the 3D images in a scrolling fashion. This choice has been made deliberately to explore the 3D classification images and see if subjects are capable of integrating multiple slices into a localization response. The result should not be interpreted as a general finding in all 3D image displays.

The images used here are based on Gaussian textures, as needed for computation of localization efficiency and the classification image technique. These images have some general similarity to anatomical variability and acquisition noise, but there are considerable differences as well including differences from smoothing filters and the ramp-spectrum of noise in tomographic imaging modalities. It may be that the results here are specific to such textures and do not extend to more realistic medical images. For example, it is possible that when image structure is present in the image, in the form of patient anatomy, it allows clinical readers to integrate across multiple slices in a way that they do not in these image stimuli. While we recognize these limitations, we also believe that this study presents baseline results that will be useful for understanding human observer performance in 3D images.

### 5 Conclusions

The main finding of this study is the limited and inefficient weighting of multiple slices in the 3D localization tasks, and the similarity of the weighting profile of the central slice to the weighting profile of the 2D tasks. The lack of integration across multiple slices provides an explanation for an observed dissociation in which large targets are more efficiently localized in the 2D tasks, and small targets are more efficiently localized in 3D tasks. This finding is consistent with the common practice of using multiple views of 3D medical images in clinical settings. The similarity between the 2D classification image and the central slice of the 3D classification image provides a rationale for using 2D tasks as a proxy for more time-consuming 3D tasks, but only under the strong assumption that other components of the search process do not disrupt this relationship.

When the observed classification images are used as a simple scanning model of localization performance, the average efficiency of the classification images is ~10% to 16% greater than the efficiency of the human subjects, which is remarkably consistent with previous findings.<sup>36</sup> However, this relationship is much weaker than previously reported ( $R^2 = 14%$  or  $R^2 = 68%$  with one outlier excluded), which indicates that other factors in the human subjects or the experimental design impact task performance.

## Disclosures

The authors have no conflicts of interest to declare.

## Acknowledgments

This work was supported by funding from the National Institutes of Health (NIH) (R01-EB026427 and R01-EB025829) and was based partly on scientific content previously reported in the SPIE Medical Imaging meeting.<sup>35</sup>

## References

1. H. H. Barrett and K. J. Myers, *Foundations of Image Science*, Wiley Series in Pure and Applied Optics, pp. xli, 1540, Wiley-Interscience, Hoboken, New Jersey (2004).
2. J. T. Bushberg et al., *The Essential Physics of Medical Imaging*, 4th ed., Wolters Kluwer, Philadelphia, Pennsylvania (2021).
3. D. Maupu et al., "3D stereo interactive medical visualization," *IEEE Comput. Graphics Appl.* **25**(5), 67–71 (2005).
4. D. J. Getty and P. J. Green, "Clinical applications for stereoscopic 3D displays," *J. Soc. Inf. Disp.* **15**(6), 377–384 (2007).
5. Z. Lu and Y. Sakamoto, "Holographic display methods for volume data: polygon-based and MIP-based methods," *Appl. Opt.* **57**(1), A142–A149 (2018).
6. P. S. Calhoun et al., "Three-dimensional volume rendering of spiral CT data: theory and method," *Radiographics* **19**(3), 745–764 (1999).
7. M. Smelyanskiy et al., "Mapping high-fidelity volume rendering for medical imaging to CPU, GPU and many-core architectures," *IEEE Trans. Visualization Comput. Graphics* **15**(6), 1563–1570 (2009).
8. G. D. Rubin et al., "Perspective volume rendering of CT and MR images: applications for endoscopic imaging," *Radiology* **199**(2), 321–330 (1996).
9. D. J. Kadrmas et al., "Impact of time-of-flight on PET tumor detection," *J. Nucl. Med.* **50**(8), 1315–1323 (2009).
10. N. J. Packard et al., "Effect of slice thickness on detectability in breast CT using a pre-whitened matched filter and simulated mass lesions," *Med. Phys.* **39**(4), 1818–1830 (2012).
11. H. W. Tseng et al., "Assessing image quality and dose reduction of a new x-ray computed tomography iterative reconstruction algorithm using model observers," *Med. Phys.* **41**(7), 071910 (2014).
12. D. Racine et al., "Task-based quantification of image quality using a model observer in abdominal CT: a multicentre study," *Eur. Radiol.* **28**(12), 5203–5210 (2018).
13. A. E. Burgess et al., "Efficiency of human visual signal discrimination," *Science* **214**(4516), 93–94 (1981).
14. A. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," *J. Opt. Soc. Am. A* **1**(8), 900–5 (1984).
15. A. E. Burgess and H. Ghandeharian, "Visual signal detection. II. Signal-location identification," *J. Opt. Soc. Am. A* **1**(8), 906–910 (1984).
16. K. J. Myers et al., "Effect of noise correlation on detectability of disk signals in medical imaging," *J. Opt. Soc. Am. A* **2**(10), 1752–1759 (1985).
17. K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Am. A* **4**(12), 2447–2457 (1987).
18. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," *J. Opt. Soc. Am. A* **9**(5), 649–658 (1992).
19. A. E. Burgess, X. Li, and C. K. Abbey, "Visual signal detectability with two noise components: anomalous masking effects," *J. Opt. Soc. Am. A* **14**(9), 2420–2442 (1997).
20. C. K. Abbey and M. P. Eckstein, "Classification images for simple detection and discrimination tasks in correlated noise," *J. Opt. Soc. Am. A* **24**(12), B110–B124 (2007).

21. A. J. Ahumada, "Putting the visual system noise back in the picture," *J. Opt. Soc. Am. A* **4**(12), 2372–2378 (1987).
22. A. Burgess and B. Colborne, "Visual signal detection. IV. Observer inconsistency," *J. Opt. Soc. Am. A* **5**(4), 617–627 (1988).
23. Z.-L. Lu and B. A. Doshier, "Characterizing human perceptual inefficiencies with equivalent internal noise," *J. Opt. Soc. Am. A* **16**(3), 764–778 (1999).
24. A. Ahumada and A. Watson, "Equivalent-noise model for contrast detection and discrimination," *J. Opt. Soc. Am. A* **2**(7), 1133–1139 (1985).
25. G. E. Legge, D. Kersten, and A. E. Burgess, "Contrast discrimination in noise," *J. Opt. Soc. Am. A* **4**(2), 391–404 (1987).
26. C. K. Abbey and M. P. Eckstein, "Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer," *J. Vision* **6**(4), 4–55 (2006).
27. C. K. Abbey and M. P. Eckstein, "Observer efficiency in free-localization tasks with correlated noise," *Front. Psychol.* **5**, 1–13 (2014).
28. C. Lartizien, P. E. Kinahan, and C. Comtat, "A lesion detection observer study comparing 2-dimensional versus fully 3-dimensional whole-body PET imaging protocols," *J. Nucl. Med.* **45**(4), 714–723 (2004).
29. J.-S. Kim et al., "A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging," *IEEE Trans. Nucl. Sci.* **51**(1), 34–40 (2004).
30. I. Reiser and R. M. Nishikawa, "Human observer performance in a single slice or a volume: effect of background correlation," *Lect. Notes Comput. Sci.* **6136**, 327–333 (2010).
31. L. Platasa et al., "Channelized Hotelling observers for the assessment of volumetric imaging data sets," *J. Opt. Soc. Am. A* **28**(6), 1145–1163 (2011).
32. C. Balta et al., "2D single-slice vs. 3D viewing of simulated tomosynthesis images of a small-scale breast tissue model," *Proc. SPIE* **10952**, 109520V (2019).
33. M. A. Lago et al., "Under-exploration of three-dimensional images leads to search errors for small salient targets," *Curr. Biol.* **31** In Press (2021).
34. P. Khurd and G. Gindi, "Decision strategies that maximize the area under the LROC curve," *IEEE Trans. Med. Imaging* **24**(12), 1626–1636 (2005).
35. C. K. Abbey, M. A. Lago, and M. P. Eckstein, "Observer templates in 2D and 3D localization tasks," *Proc. SPIE* **10577**, 105770T (2018).
36. C. K. Abbey et al., "Classification images for localization performance in ramp-spectrum noise," *Med. Phys.* **45**(5), 1970–1984 (2018).
37. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Med. Phys.* **28**(4), 419–437 (2001).
38. K. G. Metheany et al., "Characterizing anatomical variability in breast CT images," *Med. Phys.* **35**(10), 4685–4694 (2008).
39. L. Chen et al., "Anatomical complexity in breast parenchyma and its implications for optimal breast imaging strategies," *Med. Phys.* **39**(3), 1435–1441 (2012).
40. E. Engstrom, I. Reiser, and R. Nishikawa, "Comparison of power spectra for tomosynthesis projections and reconstructed images," *Med. Phys.* **36**(5), 1753–1758 (2009).
41. H. Onishi et al., "Phantom study of in-stent restenosis at high-spatial-resolution CT," *Radiology* **289**(1), 255–260 (2018).
42. L. J. Oostveen et al., "Physical evaluation of an ultra-high-resolution CT scanner," *Eur. Radiol.* **30**, 2552–2560 (2020).
43. A. M. Hernandez et al., "Validation of synthesized normal-resolution image data generated from high-resolution acquisitions on a commercial CT scanner," *Med. Phys.* **47**(10), 4775–4785 (2020).
44. M. A. García-Pérez, "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," *Vision Res.* **38**(12), 1861–1881 (1998).
45. R. F. Wagner and G. G. Brown, "Unified SNR analysis of medical imaging systems," *Phys. Med. Biol.* **30**(6), 489–518 (1985).
46. A. B. Watson and D. G. Pelli, "QUEST: a Bayesian adaptive psychometric method," *Percept. Psychophys.* **33**(2), 113–120 (1983).



47. S. A. Klein, "Measuring, estimating, and understanding the psychometric function: a commentary," *Percept. Psychophys.* **63**(8), 1421–1455 (2001).
48. D. G. Pelli, "Uncertainty explains many aspects of visual contrast detection and discrimination," *J. Opt. Soc. Am. A* **2**(9), 1508–1532 (1985).
49. D. Kersten, "Statistical efficiency for the detection of visual noise," *Vision Res.* **27**(6), 1029–1040 (1987).
50. D. Kersten, "Spatial summation in visual noise," *Vision Res.* **24**(12), 1977–1990 (1984).
51. A. J. Ahumada and J. Lovell, "Stimulus features in signal detection," *J. Acoust. Soc. Am.* **49**(6B), 1751–1756 (1971).
52. R. F. Murray, "Classification images: a review," *J. Vision* **11**(5), 2 (2011).
53. C. K. Abbey and M. P. Eckstein, "Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments," *J. Vision* **2**(1), 5 (2002).
54. R. F. Murray, P. J. Bennett, and A. B. Sekuler, "Optimal methods for calculating classification images: weighted sums," *J. Vision* **2**(1), 6 (2002).
55. C. K. Abbey and M. P. Eckstein, "Optimal shifted estimates of human-observer templates in two-alternative forced-choice experiments," *IEEE Trans. Med. Imaging* **21**(5), 429–440 (2002).
56. C. K. Abbey et al., "Approximate maximum likelihood estimation of scanning observer templates," *Proc. SPIE* **9416**, 94160O (2015).
57. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B* **57**(1), 289–300 (1995).
58. M. P. Eckstein, M. A. Lago, and C. K. Abbey, "The role of extra-foveal processing in 3D imaging," *Proc. SPIE* **10136**, 101360E (2017).
59. M. A. Lago et al., "Interactions of lesion detectability and size across single-slice DBT and 3D DBT," *Proc. SPIE* **10577**, 105770X (2018).
60. M. A. Lago et al., "Measurement of the useful field of view for single slices of different imaging modalities and targets," *J. Med. Imaging* **7**(2), 022411 (2020).
61. L. A. Remington, *Clinical Anatomy and Physiology of the Visual System*, 3rd ed., Elsevier Health Sciences, St. Louis, Missouri (2011).
62. R. G. Swensson and P. F. Judy, "Detection of noisy visual targets: models for the effects of spatial uncertainty and signal-to-noise ratio," *Percept. Psychophys.* **29**(6), 521–534 (1981).
63. R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," *Med. Phys.* **23**(10), 1709–1725 (1996).
64. H. C. Gifford et al., "A comparison of human and model observers in multislice LROC studies," *IEEE Trans. Med. Imaging* **24**(2), 160–169 (2005).
65. H. C. Gifford, Z. Liang, and M. Das, "Visual-search observers for assessing tomographic x-ray image quality," *Med. Phys.* **43**(3), 1563–1575 (2016).

**Craig K. Abbey** is a researcher in the Department of Psychological & Brain Sciences at UC Santa Barbara. His training is in the field of applied mathematics, and his research focuses on assessment of medical imaging devices and image processing in terms of performance in diagnostic and quantitative tasks.

**Miguel A. Lago** was a postdoctoral scholar in the Department of Psychological and Brain Sciences at the University of California Santa Barbara. He has recently moved to the Food and Drug Administration as a visiting scientist. His background is in computer engineering and his research studies how visual search in 3D medical image modalities affect observer performance and efficiency in radiology.

**Miguel P. Eckstein** is a professor in the Department of Psychological and Brain Sciences and affiliate faculty in the Department of Electrical and Computer Engineering at the University of California Santa Barbara. His research uses a variety of tools including behavioral psychophysics, eye tracking, electro-encephalography, functional magnetic resonance imaging, and computational modeling to study how humans see. His findings are applied to problems in medical imaging, computer vision, and interactions between robots/computer systems and humans.