Aided target recognition visual design impacts on cognition in simulated augmented reality

Aaron L. Gardony[1,2]*, Kana Okano[2], Gregory I. Hughes[1,2], Alex J. Kim[2], Kai T. Renshaw[3] and Aldis Sipolins[4]

[1]U.S. Army Combat Capabilities Development Command (DEVCOM) Soldier Center, Natick, MA, United States, [2]Center for Applied Brain and Cognitive Sciences (CABCS), Medford, MA, United States, [3]Tufts University, Medford, MA, United States, [4]Draper, Cambridge, MA, United States

Aided target recognition (AiTR) systems, implemented in head-mounted and in-vehicle augmented reality (AR) displays, can enhance human performance in military operations. However, the visual appearance and delivery of AiTR may impact other important critical aspects of human performance like decision making and situational awareness (SA). Previous research suggests salient visual AR cueing, such as found in Computer-Aided Detection diagnostic systems, orient attention strongly toward cued targets leading to missed uncued targets, an effect which may be lessened by providing analog information about classification uncertainty and using less visually salient cueing techniques, such as soft highlighting. The objective of this research was to quantify the human performance impacts of two different types of AR AiTR visualizations in a simulated virtual reality defensive security task. Participants engaged in a visual camouflage discrimination task and a secondary SA Task in which participants observed and reported a peripheral human target. Critically, we manipulated the type of AiTR visualization used: 1) a traditional salient bounding box, 2) a softly glowing soft highlight, and 3) a baseline no-AiTR condition. Results revealed minimal impacts of the visual appearance of AiTR on target acquisition, target categorization, and SA but an observable reduction in user experience associated with soft highlight AiTR. Future research is needed to explore novel AiTR designs that effectively cue attention, intuitively and interpretably visualize uncertainty, and deliver acceptable user experience.

KEYWORDS

aided target recognition, augmented reality, virtual reality, target acquisition, decision making, situational awareness, eye tracking

1 Introduction

The United States Army intends to leverage advances in artificial intelligence and machine learning (AI/ML) in future warfighter systems, including its combat vehicles and augmented reality (AR) displays (Army.mil, 2022; Larkin et al., 2019; Sam. gov, 2022). For emerging head-mounted AR displays, AI/ML can enable a decisive edge to rapidly detect and classify potential threats. Such Aided Target Recognition, or AiTR, technologies use computer vision algorithms, such as the widely used You Only Look Once class of object detection algorithms (Sultana et al., 2020), to detect, label, and highlight potential targets. In military settings, AiTR can assist with a myriad of mission functions including fire control, surveillance, reconnaissance, intelligence, persistent surveillance, and situational awareness (Ratches, 2011). The speed at which AiTR algorithms can detect and highlight potential targets is a significant asset to Soldiers battling in the kinetic, fast-paced multidomain operations of the future (US Army TRADOC, 2018). However, improperly designed AiTR may interfere with other aspects of task performance, leading to increased false alarms, attentional tunneling, and degraded situational awareness (SA) (Geuss et al., 2019). The present study addresses these issues, investigating the cognitive impacts of AiTR through quantitative assessments of human performance in a virtual reality (VR) lethal force decision making scenario incorporating simulated AiTR AR overlays.

Current implementations of AiTR typically highlight potential targets of interest via opaque, colored bounding boxes. This highly salient display readily captures human attention, serving the main purpose of AiTR well. However, such salient cues can lead to attentional tunneling, drawing attention away from the presence of unexpected targets (Yeh & Wickens, 2001; Wickens, 2005; Chen & Barnes, 2008). Moreover, in the medical domain, such as in the fields of surgery and radiology, researchers have found Computer-Aided Detection diagnostic systems that use salient AR cueing can draw attention away from other important information during visual search or lead diagnosticians to prematurely cease active searching, especially under conditions of low target prevalence (Krupinski et al., 1993; Philpotts, 2009; Drew et al., 2012, 2020; Cain et al., 2013; Dixon et al., 2013). Such attentional cueing effects can have grave consequences for Soldier users of AiTR systems, reducing their SA for non-cued features in their environments, such as potential threats (Endsley, 1995). To address attentional tunneling issues associated with salient AR cueing, researchers have developed a variety of subtle cueing techniques that modulate background luminance, color temperature, contrast, and opacity with demonstrated benefits in target identification (McNamara et al., 2008; Bailey et al., 2009; Lu et al., 2012, 2014).

One such technique is soft highlighting, which cues areas of interest with a soft blur. Kneusel and Mozer's (2017) initial experiments testing soft highlighting demonstrated consistent visual search performance advantages over hard highlighting. Given these findings, soft highlighting could be used in AR systems, as proposed by Larkin et al. (2019), to display AiTR cues in a non-distracting manner. Moreover, in contrast with the salient hard highlighting of bounding boxes, soft highlighting may be preferable in situations with multiple overlapping highlighted targets, distinguishing probable targets from the environment while reducing visual clutter and missed uncued targets (Fenton et al., 2007; 2011). Notably, subtle AR cueing techniques, like soft highlighting, have been developed and validated in low-stakes visual search contexts rather than the high-stakes contexts in which AiTR systems are intended to be used (e.g., military operations). Moreover, extant research examining human performance impacts of AiTR has primarily examined visual search of static images rather than dynamic scenes involving the type of tasks and decision making military operators using future AiTR systems would potentially encounter.

In the present study we investigated the human performance impact of simulated AiTR AR cues in a VR Lethal Force Decision Making (LFDM) task, focusing on three performance areas: target acquisition, target categorization, and SA. Participants categorized a single animated Soldier avatar advancing towards them amongst non-combatant civilians as friendly or enemy based upon their worn camouflage pattern, deciding to shoot them or let them pass using handheld VR controllers. Participants completed the LFDM task in three blocks, two with AR AiTR overlays (bounding box, soft highlight) and one without. As a first step we rendered AiTR without incorporating uncertainty information but subtly varied AiTR appearance and included incorrect civilian overlays (i.e., false alarms) to reflect limitations of current AiTR systems. We also incorporated a simultaneous secondary detection task ("SA Task") in which participants responded when they observed a non-combatant civilian in their periphery. We measured multiple quantitative metrics, including behavioral responses (e.g., accuracy, response time) and eye tracking metrics (e.g., dwell time, glance latency) to assess how AiTR impacted performance in the LFDM and SA Tasks.

2 Materials and methods

2.1 Participants

Forty-one male active-duty Soldiers (M_{age} = 22.2, SD_{age} = 3) voluntarily participated. Human use approvals were reviewed and approved by the United States Army Combat Capabilities Development Command Soldier Center Human Research Protection Program Office and the Tufts University Institutional Review Board. Written informed consent for participation was not required for this study in accordance

with the national legislation and the institutional requirements. This study was conducted during the SARS-CoV-2 pandemic. All researchers and participants followed safety protocols approved by U.S. Army and Tufts University safety committees (e.g., masks, sanitization). All participants possessed 20/25 binocular distance visual acuity or better as determined by the Snellen eye chart and normal color vision as determined by the Color Vision Testing Made Easy test (Cotter et al., 1999). One participant was removed from analysis due to an error in data collection. Forty participants (M_{age} = 21.76, SD_{age} = 2.86) were ultimately included in the analysis.

2.2 Materials

2.2.1 Apparatus

The VR LFDM task was developed using Unity3D software (Unity Technologies, 2021). The task was rendered and displayed using SteamVR and Varjo Base software (Valve Corporation, 2021; Varjo, 2021) on the high resolution Varjo VR-3 head-mounted display (HMD) which has a display resolution of 70 pixels-per-degree, a 90 Hz display refresh rate, a 115° horizontal field of view, and advertised sub-degree 200 Hz eye tracking. Participants were seated during the experiment and participant responses were tracked using HTC Vive VR controllers (HTC, 2021). The experiment was run on a Lenovo ThinkStation P920 Workstation with an Intel Xeon Gold 6,246 processor and NVIDIA Quadro RTX 8000 dedicated graphics card. Behavioral and eye tracking data were logged via custom Unity scripts and Varjo's Unity XR SDK.

2.2.2 Virtual reality scenario
2.2.2.1 Overall scenario design and trial structure

The LFDM task was designed to evaluate the impact of AiTR overlays on target acquisition, target categorization, and SA in a Soldier-relevant task. The task approximated a defensive security task (DST) in which the participant is stationed on patrol at an Entry Control Point and tasked to defend it from potential threats. Threats took the form of a Soldier advancing toward the Entry Control Point and participants had to make lethal force decisions about whether to shoot the Soldier or let them pass. This task captured metrics related to target acquisition and categorization. A simultaneous secondary task ("SA Task") involved "radioing in" non-combatant civilians who appeared in nearby windows in the periphery. This task captured metrics related to SA. Below, we describe implementation details for each embedded task, in turn.

The DST was inspired by similar research tasks that probe decision making under conditions of perceptual uncertainty (Brunyé & Gardony, 2017; Brunyé and Giles, under review). In this task, participants distinguish between friendly and enemy versions of a camouflage pattern worn by a walking Soldier avatar. We developed two distinguishable grayscale camouflage

patterns using a binary space partitioning algorithm (Åström, 2011) and then introduced uncertainty by layering them and systematically altering the top layer opacity in photo editing software. This procedure resulted in six blended patterns with varying uncertainty: 100% friendly, 65% friendly, 51% friendly, 51% enemy, 65% enemy, and 100% enemy. Each pattern had an objectively correct answer (i.e., ground truth). We then used Unity3D to texture a Soldier avatar with each pattern. Figure 1 shows an example trial of the DST (A) and example Soldier avatars with overlayed camouflage patterns (B).

Concurrently with the DST, the SA Task involved participants monitoring for the appearance of a single non-combatant civilian ("SA Target") in the windows of nearby buildings. On each trial the SA Target could appear randomly in one of four windows; two located in central vision (12° and 14° visual angle) and two located in near-peripheral vision (25° and 31° visual angle). After appearing, the SA Target would continuously and randomly move slightly in the window to attract attention. Figure 1A shows an example SA Target and the four possible spawn locations.

A LFDM trial consisted of a single Soldier avatar emerging from one of four locations adjacent to the scenario's central corridor and walking toward the participant at a brisk walking pace. The near-distance and far-distance Soldier spawn locations were positioned approximately 50 m (approximately 16 s from spawn position to red line) and 100 m (approximately 22 s from spawn position to red line), respectively, from the participant on either side of the central corridor. To prevent predictable stimulus timing, the Soldier spawned after a random delay ranging from 1 to 5s following the start of the trial. Similarly, the SA Target's spawn times were randomly delayed based on the Soldier spawn location: the SA Target spawned 3–12 s following a near-distance Soldier spawn and 3 – 18 s following a far-distance Soldier spawn. Notably, 50% of the time, the SA Target would not spawn unless the random delay had elapsed *and* the participant was currently gazing at the Soldier. Such *gaze-contingent* SA Target spawns were designed to encourage attentional focus on the DST and to prevent visual scanning strategies of the adjacent buildings and windows. Additionally, throughout the trial 15 civilian non-player characters walked around the environment, in and out of view, to provide an element of realism and visual clutter. This resulted in ~10 civilians being in view at any given time on average. Civilian movements were restricted to areas outside the central corridor to prevent significant overlap with the advancing Soldier. The trial ended once the Soldier was categorized as friendly or enemy or if the Soldier crossed the red line directly in front of the participant, which marked the trial as a non-response. Practically, this meant that a trial could end prior to the SA Target appearing.

2.2.2.2 AiTR augmented reality overlays

We developed two simulated AR AiTR overlays, a bounding box and a soft highlight overlay, leveraging a commercially-
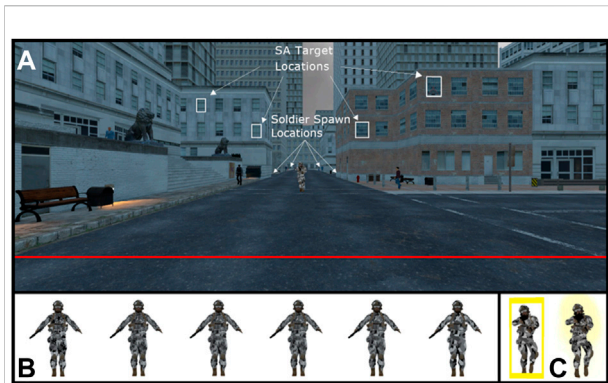
FIGURE 1
(A) Example trial of the Defensive Security Task (DST) from the No AiTR condition. We have added arrows depicting the four possible Soldier spawn locations and the four possible SA Target spawn locations. Example SA Target is shown in top-right window. **(B)** Example Soldier avatars in the six possible blended camouflage patterns. From left to right, 100% Enemy, 65% Enemy, 51% Enemy, 51% Friendly, 65% Friendly, 100% Friendly. **(C)** Example AR AiTR overlays. Left side depicts Bounding Box overlay (BB). Right side depicts Soft Highlight overlay (SH).

available Unity asset package (Kronnect, 2021), depicted in Figure 1C. Consistent with previous research demonstrating its utility, we rendered AiTR overlays in yellow (Tombu et al., 2016; Reiner et al., 2017; Larkin et al., 2019). Bounding boxes surrounded the Soldier without overlaying their body or the camouflage pattern. Similarly, the soft highlight emanated from the Soldier without overlaying their body or the camouflage pattern. To approximate current technological limitations, we introduced random dynamic jitter to the AiTR display, subtly shifting the X and Y scale of the bounding box overlays and the "glow width" parameter of the soft highlight every 0.25s. We also introduced error in the form of false alarm AiTR. On any given trial, each civilian had a one third chance of being assigned an AiTR overlay. Practically, this resulted in three to four civilians in view being highlighted at any given time on average. This design choice was intended to increase visual clutter and reflect inconsistency of the AiTR accurately extracting the Soldier from the civilians.

2.3 Factors and performance measures

The LFDM task manipulated several factors of interest. The advancing Soldier was either friendly or enemy (Soldier Type: Friendly, Enemy) and was wearing one of three camouflage pattern mixtures which influenced the underlying uncertainty of the categorization process (Camo Clarity: 51, 65, 100%). Across blocks, the AiTR was manipulated such that the advancing Soldier (and "false alarm" civilians) either had No AiTR overlays, a Bounding Box overlay, or a Soft Highlight overlay.

The LFDM task recorded several performance measures across the embedded DST and SA Tasks, encompassing participants' behavioral response (i.e., button responses) and measures of visual attention derived from eye gaze. For the DST task, the behavioral dependent variables were Categorization Error of the advancing Soldier (i.e., the proportional rate of error responses in the friend-foe categorization task) and Categorization Distance (i.e., how far away the Soldier was in meters from the participant at response). Soldier Glance Latency was recorded as the latency (in seconds) from when the advancing Soldier first appeared in view and when the participant glanced at the Soldier. We also recorded Relative Dwell Time as the proportional amount of time the participant's gaze point overlayed or "dwelled" on various features of the VR scenario, including the advancing Soldier, the adjacent buildings' windows, the civilians, and the general environment (e.g., roads, buildings, sky, etc.). It should be noted that the accuracy of gaze-based measures is influenced by the distance of the gaze target; further gaze targets subtend a smaller visual angle leading to reduced accuracy. To address this, the task implemented a dynamically sized gaze intersection zone surrounding the advancing Soldier, SA Target, and civilians that continuously subtended 2° visual angle (vertical and horizontal). For the SA Task, we measured 1) P(SA Response), the response rate of the SA Target, 2) Task Response Time, 3) P(SA Glance), the rate at which the participant glanced at the SA Target, given the target appeared during the trial, and 4) SA Target Glance Latency, the latency (in seconds) from when the SA Target first appeared in the windows and when the participant glanced at the SA Target.

2.4 Procedure

Data were collected in a single 2-h experimental session. We used a within-participants design with participants completing the LFDM task across three AiTR blocks, the order of which was counterbalanced across participants. At the start of the session, the researcher instructed participants on how to don the HMD, use the controllers, and briefly explained the task. Detailed task instructions were given later in the HMD. The researcher explained that participants would engage in a defensive security task in which they were tasked with identifying whether an approaching Soldier, embedded amongst non-combatant civilians, was friendly or enemy based on the type of camouflage pattern on their uniform. In the two AiTR blocks, advancing Soldiers were highlighted in yellow. Participants were told that AiTR technology is not perfect and that it may highlight civilians who are not wearing camouflage. They were asked to ignore these civilians and only focus on the advancing Soldier. No highlighting was present on either the Soldier or civilians in the block without AiTR. Participants marked Friendly Soldiers with a trigger button on the Vive controller in their

left hand and Enemy Soldiers with the trigger button in their right hand. Participants were also tasked to observe and report the appearance of a civilian in the windows (SA Task), pressing the left trackpad when they detected a civilian in a left window and the right trackpad for a civilian detected in a right window. Each AiTR block comprised 24 randomized-order trials that fully combined four Soldier spawn locations near-distance (left/right), far-distance (left/right) and six Camo Clarity levels (100% friendly, 65% friendly, 51% friendly, 51% enemy, 66% enemy, 100% enemy).

Prior to completing each AiTR block, participants completed a VR sickness questionnaire (Kim et al., 2018) and donned the HMD. 17-point calibration and validation of the eye tracker was run to ensure accurate eye tracking ($M_{error} < 2°$). Participants then studied the 100% friendly and enemy camouflage patterns for 2 minutes and took a pattern identification test to ensure they could accurately distinguish the friendly pattern from the enemy. To pass this test, participants had to correctly identify 10 camouflage patterns. Participants also completed five practice trials before the start of each condition to ensure they understood the task. At the end of each AiTR block, participants completed the VR sickness questionnaire (Kim et al., 2018) again as well as the NASA Task Load Index (NASA-TLX) cognitive workload assessment (Hart & Staveland, 1988) to assess perceived workload associated with each AiTR condition. Following completion of the three blocks, the researcher debriefed the participant on the purpose of the study and answered any questions.

2.5 Analysis approach

To analyze the data, we employed linear mixed models (LMMs). We used R (R Core Team, 2022) and the afex package (Singmann et al., 2020) to run the LMMs. We approximated degrees of freedom (df) and computed p values via the Satterthwaite method unless otherwise noted. We conducted follow-up pairwise comparisons (with Tukey-adjusted p values) with the emmeans package (Lenth, 2020) and visualized the data with ggplot2 (Wickham, 2016). We expected individuals would vary in their baseline performance and we included by-participants random intercepts across our analyses, e.g., model = DV ~ Factor1 * Factor 2. . . + (1| Participant), unless otherwise noted.

3 Results

3.1 Defensive security task results

Prior to analysis we first removed non-response trials where the advancing Soldier crossed the red line (<1%). Next, we removed trials where the participant never glanced at the advancing Soldier or if Soldier Glance Latency exceeded 5 s. If

this procedure resulted in greater than 50% of trials being discarded for a participant's block, we discarded the entirety of the data for that block. This procedure, which was intended to omit poor quality data from experimental blocks with technical difficulties and/or high levels of participant drowsiness, removed 12% of trials. Percentages of removed trials did not significantly differ across the three AiTR conditions. Removed trials did not differ in any systematic way across our manipulated factors of interest.

3.1.1 Behavioral responses

We first scrutinized Categorization Error in the DST, submitting these data (binomial responses: 0, 1) to a 2(Soldier Type) x 3 (AiTR) x 3(Camo Clarity) repeated measures binomial generalized LMM and computed df and p values using likelihood ratio tests (LRTs). Participants were biased overall to categorize the advancing Soldier as enemy ($M_E = 58\%$, $M_F = 42\%$). Observed errors were lowest in the 100% Camo Clarity (100CC) condition ($M = 6.6\%$) followed by the 65CC ($M = 12.2\%$) and 51CC ($M = 50\%$) conditions, $X (2, n = 40) = 499.17$, $p < 0.001$, and were lower for Enemy ($M = 15\%$) vs. Friendly ($M = 31\%$) Soldiers, $X (1, n = 40) = 19.88$, $p < 0.001$. These main effects were qualified by a Camo Clarity x Soldier Type interaction, $X (2, n = 40) = 59.93$, $p < 0.001$. As shown in Figure 2A, errors were significantly higher for Friendly Soldiers relative to Enemy in the 51CC and 65CC conditions, p's < 0.001, but this difference reversed in the 100CC condition, $p < 0.05$. We also observed a Camo Clarity × AiTR interaction, $X (4, n = 40) = 12.21$, $p < 0.05$, however no notable pairwise comparisons emerged significant making the interaction difficult to interpret so we do not discuss it further. No other AiTR effects or others emerged significant.

Next, we investigated the Categorization Distance data, submitting these data to a 2 (Soldier Type) x 3 (AiTR) x 3 (Camo Clarity) repeated measures LMM. Soldier distance at participant response was furthest in the 100CC condition ($M = 32.9$ m) followed by the 65CC ($M = 25$ m) and 51CC ($M = 22$ m) conditions, $F (2, 2,479.2) = 128.75$, $p < 0.001$, and was further for Friendly ($M = 28.7$ m) vs. Enemy Soldiers ($M = 24.6$ m), $F (1, 2,479.2) = 48.83$, $p < 0.001$. These main effects were qualified by a Camo Clarity x Soldier Type interaction, $F (2, 2,479.2) = 30.32$, $p < 0.001$. As shown in Figure 2B, the differences in Categorization Distance for Friendly and Enemy Soldiers were driven by 100CC trials, $p < 0.001$. We also observed a significant AiTR main effect, $F (2, 2,486.3) = 4.31$, $p < 0.05$. Categorization Distances in the Soft Highlight ($M = 27.3$ m) and No AiTR ($M = 27.7$ m) conditions were significantly longer than the Bounding Box condition ($M = 25$ m), and did not significantly differ from each other.

3.1.2 Eye gaze

Next, we scrutinized eye gaze movement in the DST. We first investigated Soldier Glance Latency to understand if the manipulated factors influenced visual acquisition of the
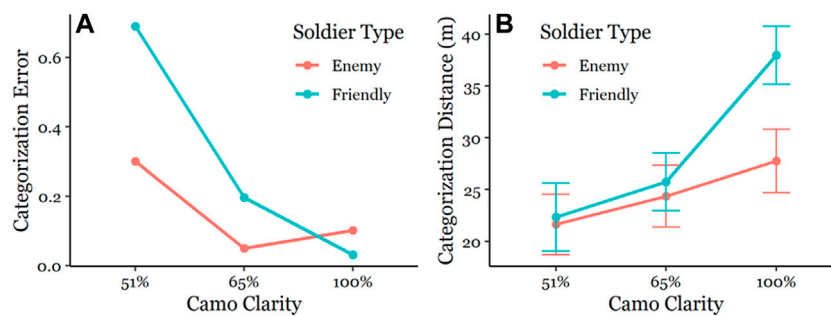
FIGURE 2
(A) Mean Categorization Error as a function of Camo Clarity and Soldier Type. Errors were significantly higher for Friendly Soldiers relative to Enemy Soldiers in the 51CC and 65CC conditions ($p < 0.001$). Errors were significantly higher for Enemy Soldiers relative to Friendly Soldiers in the 100CC condition ($p < 0.001$). **(B)** Mean Categorization Distance (m) as a function of Camo Clarity and Soldier Type. Categorization Distance for Friendly vs. Enemy Soldiers were significantly different in the 100CC condition ($p < 0.001$). Error bars depict standard error of the mean.

advancing Soldier. We submitted these data to a 2(Soldier Type) x 3 (AiTR) x 3(Camo Clarity) repeated measures LMM. Latencies were fastest in the 100CC condition ($M = 0.68$s) followed by the 65CC ($M = 0.74$s) and 51CC ($M = 0.86$s) conditions, F (2, 2,481.1 = 7.78), $p < 0.001$. Pairwise comparisons revealed 100 and 65CC latencies were significantly faster than 51CC ($p < 0.001$, $p < 0.05$, respectively) but did not differ from each other. No AiTR effects or others emerged significant.

Second, we investigated Relative Dwell Time to the scenario's environmental features. We first calculated the intersection point of the participant's gaze vector to relevant scenario features: 1) the Soldier, 2) the civilians, 3) the Windows, 4) the SA Target, and 5) the Environment (i.e., every scene element not falling into one of the other categories) and computed the by-trial total duration that participants' gaze intersected with each feature. Dwell time toward the SA Target was very low ($M = 0.15$ s) because participants rapidly responded upon seeing the SA Target so we removed the SA Target from the Relative Dwell Time analysis. We then computed a proportional Relative Dwell Time metric by dividing by-trial dwell time for each scenario feature by the total trial dwell time (excluding SA Target dwell times). We then submitted these data to a 4 (Scenario Feature) x 2 (Soldier Type) x 3 (AiTR) x 3 (Camo Clarity) repeated measures LMM. We expected Relative Dwell Times across scenario features would vary between individuals in addition to varying baseline performance and so we included random slopes for Scenario Feature as well as by-participants random intercepts.

Relative Dwell Time differed across scenario features; on average, participants dwelled on the Soldier the longest ($M = 55\%$ of the time), followed by the Windows ($M = 21\%$), the Environment ($M = 13\%$), and the civilians ($M = 11\%$), F (3, 38.52) = 269.11, $p < 0.001$ (all pairwise comparisons $p < 0.001$). This main effect was qualified by a Scenario Feature x Soldier Type interaction, F (3, 9,955.97) = 4.04, $p < 0.01$, a Scenario
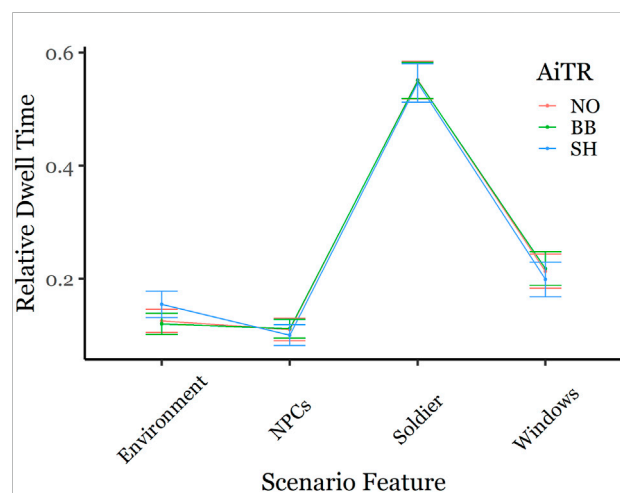

FIGURE 3
Mean Relative Dwell Time on LFDM Scenario Features across AiTR conditions (NO: No AiTR, BB: Bounding Box, SH: Soft Highlight). SH significantly increased Relative Dwell Time toward the environment compared to NO and BB conditions ($p < 0.001$). Error bars depict standard error of the mean.

Feature × AiTR interaction, F(6, 9,961.87) = 5.52, $p < 0.001$, and a Scenario Feature x Soldier Type x Camo Clarity interaction, F (3, 9,955.6) = 2.56, $p < 0.05$. As shown in Figures 3, 4, the observed interactions suggest that 1) Soft Highlight AiTR significantly increased Relative Dwell Time toward the environment at the expense of other scenario features compared to the No AiTR and Bounding Box conditions, $p < 0.001$ (Figure 3), and 2) participants dwelled on the advancing enemy Soldiers relatively longer than Friendly Soldiers, most apparently when they wore the clearest camo
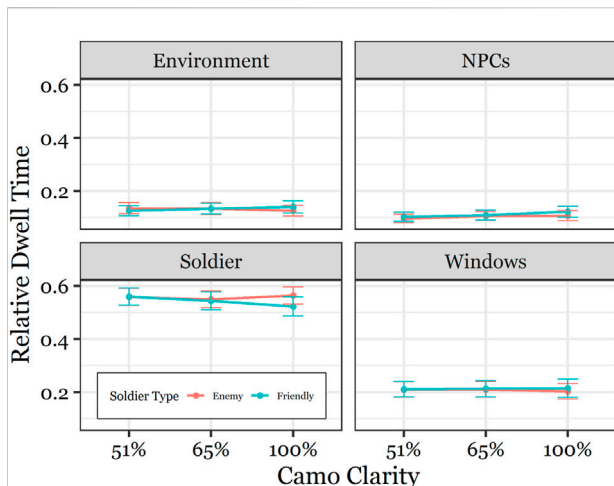
FIGURE 4
Mean Relative Dwell Time on LFDM Scenario Features as a function of Camo Clarity and Soldier Type. Participants dwelled on the advancing enemy Soldiers relatively longer than Friendly Soldiers, when they wore the clearest camo pattern (100CC) ($p < 0.01$). Error bars depict standard error of the mean.

pattern (100CC), $p < 0.01$ (Figure 4). Taken together with the Categorization Distance findings, this suggests that the shorter Enemy Categorization Distances observed in the 100CC condition can be plausibly explained by differences in visual interrogation of the advancing Soldier.

3.2 Situational awareness task results

Prior to analysis, we removed trials where SA Targets never appeared. This resulted in the removal of an additional 20% of the data. Recall that on each trial a single SA Target appeared in one of four window locations. SA Target spawns were either Timed or Gaze-Contingent, the appearance of which depended on the participant dwelling on the advancing Soldier.

Overall, when the SA Target appeared, participants successfully responded 87.5% of the time ($M_T = 89\%$, $M_{GC} = 86\%$) and glanced at it 58% of the time ($M_T = 61\%$, $M_{GC} = 56\%$). Notably, P(SA Response) and P(SA Glance) were significantly associated, X (3, $n = 40$) = 299.58, $p < 0.001$. When participants glanced at the SA Target during a trial, they responded 98% of the time. However, when they did not glance at it, they still successfully responded 72% of the time, suggesting the SA Target was reliably detectable in peripheral vision. On average, participants responded to the SA Target 1.33s ($M_T = 1.37$ s, $M_{GC} = 1.3$ s) after stimulus onset and glanced at it 0.85s ($M_T = 0.88$ s, $M_{GC} = 0.82$ s) after stimulus onset.
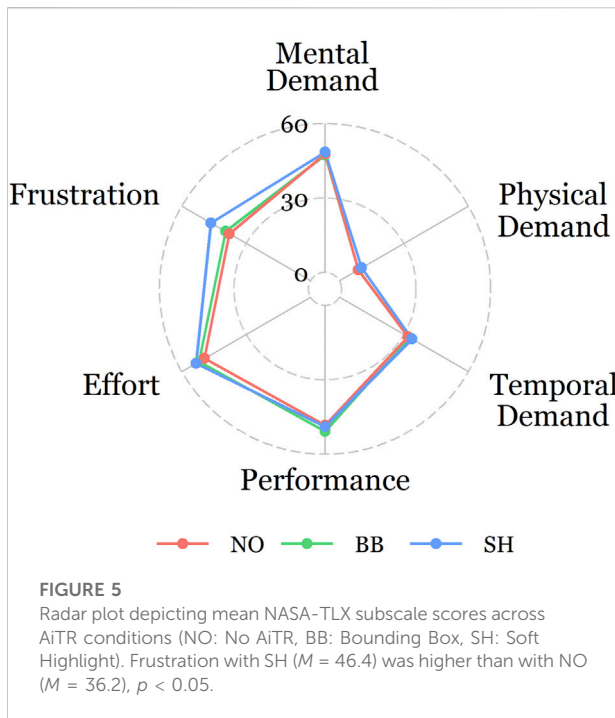
We first separately examined P(SA Response) and P(SA Glance) of the SA Target (binomial responses:0,1), submitting both to a 2(SA Target Spawn Type) x 2 (Soldier Type) x 3 (AiTR)

x 3 (Camo Clarity) repeated measures binomial LMMs and computed df and p values using LRTs. Gaze-Contingent SA Targets were responded to and glanced at ($M_{Response} = 86\%$, $M_{Glance} = 56\%$) less often than Timed targets ($M_{Response} = 89\%$, $M_{Glance} = 61\%$), P(SA Response) - X (1, $n = 40$) = 5, $p < 0.05$, P(SA Glance) - X (1, $n = 40$) = 4, $p < 0.05$. The P(SA Response) main effect was qualified by a SA Target Spawn Type × AiTR interaction, X (2, $n = 40$) = 5.98, $p = 0.05$. Pairwise comparisons revealed that while P(SA Response) was numerically lower for Gaze-Contingent vs. Timed spawns across AiTR conditions, the SA Target Spawn Type main effect appeared to be driven by differences in the Soft Highlight AiTR condition ($M_T = 90\%$, $M_{GC} = 81\%$), $p < 0.05$. We also observed a Soldier Type x AiTR x Camo Clarity interaction, X (4, $n = 40$) = 9.58, $p < 0.05$, however no pairwise comparisons emerged significant making the interaction difficult to interpret so we do not discuss it further. No other significant effects emerged. We next examined Detection Response Time of the SA Target and Glance Latency to the SA Target, submitting both to a 2(SA Target Spawn Type) x 2 (Soldier Type) x 3 (AiTR) x 3 (Camo Clarity) repeated measures LMM. No significant effects emerged.

3.3 Virtual reality sickness and cognitive workload results

Recall that before each block participants completed a pre-block VRSQ and after each block completed a post-block VRSQ and the NASA-TLX survey to provide an assessment of changes in VR sickness and cognitive workload associated with each AiTR block. We first computed the oculomotor discomfort and disorientation subscales of the VRSQ and the individually submitted subscale data to a 2(Time: Pre vs Post) x 3 (AiTR) repeated measures LMM with random slopes for Time as well as by-participant random intercepts. Oculomotor discomfort but not disorientation significantly increased following each block, F (1, 39) = 8.47, $p < 0.01$. No other significant effects emerged. We next examined the NASA-TLX scores. One participant did not complete the NASA-TLX survey after all three AiTR blocks, one after the Soft Highlight block, and one after the No AiTR block. These data were not included in the analysis. The NASA-TLX survey results are plotted in Figure 5.

We first computed the six NASA-TLX subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration) across the three AiTR conditions and then individually submitted subscale data to repeated measures LMMs with AiTR as the sole fixed effect and by-participant random intercepts. Frustration scores significantly differed across AiTR conditions, F (2, 74.33) = 4.21, $p < 0.05$. Frustration with Soft Highlight AiTR ($M = 45$) was higher than with No AiTR ($M = 36.2$), $p < 0.05$. Bounding Box AiTR ($M = 38.3$) did not significantly differ from No AiTR and Soft Highlight AiTR.

FIGURE 5
Radar plot depicting mean NASA-TLX subscale scores across AiTR conditions (NO: No AiTR, BB: Bounding Box, SH: Soft Highlight). Frustration with SH (M = 46.4) was higher than with NO (M = 36.2), p < 0.05.

All other NASA-TLX subscales did not differ across the three AiTR conditions (all ps > 0.05).

Given that Soft Highlight AiTR increased frustration, we next investigated whether this frustration related to performance in the LFDM task. We found a significant positive correlation between Frustration scores and Categorization Error in the Soft Highlight AiTR condition (r = 0.36, t (33) = 2.25, p < 0.05), suggesting that the more frustrated participants were using Soft Highlight AiTR, the more error-prone they were when categorizing the advancing Soldier as friendly or enemy. Frustration scores were not correlated with any of the other performance metrics.

4 Discussion

The present study investigated the impact of simulated Aided Target Recognition (AiTR) Augmented Reality (AR) cues on human performance in a Lethal Force Decision Making (LFDM) task. Participants categorized advancing Soldier avatars as friendly or enemy based on their worn camouflage pattern that varied in perceptual discriminability while simultaneously monitoring for a peripheral non-combatant civilian. Participants completed three different AiTR sessions: one without any AiTR overlay, one with a bounding box, and one with a soft highlight. Our primary interest was to investigate how alternative soft highlight AiTR techniques, which offer the potential to intuitively display visual classification confidence or threat classification (Larkin et al., 2019), differ from traditional

bounding boxes with respect to target acquisition, target categorization, and situational awareness (SA).

A well-established benefit of AiTR is enhanced target detectability and acquisition (Ratches, 2011). We operationalized target acquisition performance with Soldier Glance Latency by investigating the delay between when the advancing Soldier first appeared in view and when the participant first glanced at the Soldier. Interestingly, we did not observe any such latency reductions for AiTR vs No AiTR. In fact, the only significant result observed was faster latency under conditions of relative certainty vs uncertainty; Glance Latency toward Soldiers with 100% (100CC) and 65% (65CC) Camo Clarity was significantly faster than 51CC, though the numerical differences were slight (<0.2s). These unexpected results may have been due to limitations in our virtual reality (VR) scenario design. The lack of an AiTR effect on target acquisition may have stemmed from the limited Soldier spawn locations (4), making it relatively predictable to detect the appearance of a Soldier. Moreover, the observed faster Soldier Glance Latency for clearer camo patterns may be attributable to their higher visual contrast than the more ambiguous 51CC mixed camo patterns. Future studies on human performance impacts of AiTR should include many more possible spawn locations to reduce stimulus predictability and should carefully consider the visual perceptibility and discriminability of stimuli used.

AiTR did not appreciably impact categorization performance in the LFDM task. Categorization Error did not differ between AiTR conditions. We did observe further Categorization Distances for Soft Highlight vs. Bounding Box AiTR, suggesting faster friend-foe discrimination. However, it should be noted that this difference was numerically slight (~2 m). These results may be partially explained by the previously discussed null AiTR target acquisition results. Consider that AiTR could improve categorization performance in two main ways: 1) directly by conveying information about target categorization and 2) indirectly though enhanced target acquisition speed, maximizing the amount of time available to make a discriminative judgment. Given that our AiTR cues did not convey information about target categorization and did not enhance target acquisition speed it unsurprising that AiTR did not impact categorization performance.

Participants demonstrated an overall tendency to categorize advancing Soldiers as enemy rather than friendly (58 vs. 42%). This trend was especially apparent in the uncertain Camo Clarity conditions (51 and 65%) where participants tended to miscategorize Friendly Soldiers as Enemy, suggesting that under conditions of relative uncertainty, participants' friend-foe decision making was biased towards perceiving the advancing Soldier as a threat. Concurrently, under conditions of relative certainty, participants allowed the advancing Soldier to get closer before making their decision. This effect was driven by the 100CC condition in which participants judged Friendly Soldiers much more rapidly compared to Enemy Soldiers. Taken together, these

results suggest a *speed-accuracy tradeoff*: under conditions of relative certainty identifying Friendly vs Enemy Soldiers is faster but less accurate.

Overall, AiTR did not result in degraded SA due to attentional tunneling, a common pitfall of Computer-Aided Detection. Throughout the LFDM task, participants engaged in a secondary SA Task in which they detected non-combatant civilian activity in their visual periphery (i.e., SA Target). Attentional tunneling in AiTR conditions could have manifested as decreased detection rates of the SA Target, decreased glance rates at the SA Target, or longer detection and glance latencies. However, we observed no notable impacts of AiTR on these measures. Average detection rates and latencies were near ceiling, 87.5% and 1.33s, respectively. The high overall levels of performance observed for the SA Task may have stemmed from limitations of the scenario design. First, the SA Target subtly but continuously moved ("jittered") after appearing which likely increased its detectability. Second, as with the advancing Soldier, the SA Target could appear in one of four locations; participants could have capitalized on this predictability. Third, SA Target spawn locations were positioned in the near periphery, ranging from 12–31° visual angle. We chose the spawn locations to fall within the field-of-view of the VR head-mounted display (HMD) to ensure SA Targets were visible. However, real-world peripheral stimuli occur at larger eccentricities (>60°), extending beyond the field-of-view capabilities of current generation VR HMDs. Fourth, and perhaps most importantly, while SA Target presentation location and timing were randomized, the SA Target could appear on every trial and in practice did frequently (80% of trials). This stimulus frequency may have shifted the SA Task from an unpredictable secondary task to a predictable dual-task. Nevertheless, while SA Target detection performance was high overall, promisingly, we observed lower detection rates for Gaze-Contingent SA Targets, which appear only if the participant is currently dwelling on the advancing Soldier, relative to Timed SA Targets, which do not consider the participant's gaze. Of note, this difference appeared to be driven by Soft Highlight AiTR, raising the possibility that soft highlighting degrades SA for peripheral targets. Given these tentative findings, future studies assessing SA impacts of AiTR should incorporate both gaze-contingent and stationary SA Targets in secondary detection tasks to accurately calibrate stimulus detectability in dynamic VR scenarios.

Despite the null results of the main phenomena of interest, additional analyses reveal useful insights regarding how to implement AiTR in head-mounted AR displays effectively. For example, an analysis of participants' eye gaze across elements in the visual scene pointed to a potential issue with our Soft Highlight AiTR implementation. Expectedly, we found that participants spent most of their time overall (55%) dwelling on the advancing Soldier compared to other scenario features. Participants also tended to dwell on Enemy Soldiers longer than

Friendly Soldiers under conditions of relative certainty. This finding aligns with the Categorization Distance findings and suggests that the relatively shorter Categorization Distances observed for 100CC Enemy vs Friendly Soldiers were due to participants visually inspecting Enemy Soldiers longer. Notably, relative to Bounding Box and No AiTR, Soft Highlight AiTR increased relative dwell time toward the environment at the expense of other scenario features. Anecdotally, several participants reported that Soft Highlight AiTR induced ocular discomfort. The Soft Highlight AiTR was rendered with dithering, a process by which random noise is intentionally applied for visual effect. Our intent was to simulate the real-world resolution limitations of current generation AR HMDs. However, the Soft Highlight AiTR's dithered, screen door-like appearance yoked to the advancing Soldier may have produced a flickering visual effect leading to the ocular discomfort reported. This discomfort in turn could explain why participants in the Soft Highlight condition spent less time dwelling at the highlighted Soldier and more toward the environment. Indeed, there is evidence of such discomfort in the NASA-TLX survey data which demonstrated that participants were more frustrated by the Soft Highlight condition relative to No AiTR. Taken together, our findings suggest that while soft highlighting may be effective at highlighting targets in static images, such as when inspecting satellite imagery or histology slides, it negatively impacted task performance and user experience in our dynamic AiTR task.

One notable aspect of our research was our unique participant pool comprised of all-male Soldiers. Although this offers valuable insight into the behaviors of active-duty Soldiers performing a Soldier-relevant task, it also limits the generalizability of the present study's findings. Previous eye tracking research investigating sex differences in attentional cueing found no observed sex differences for exogenous cueing (Bayliss et al., 2005). Consequently, we would not expect sex differences based on AiTR in LFDM task performance. However, sex differences in violent and aggressive behavior (Staniloiu & Markowitsch, 2012) may suggest the possibility of behavioral differences in the LFDM task. Thus, future work should investigate if biological sex interacts with AiTR. We also elected to incorporate current limitations of real-world AiTR systems, such as false alarms, rather than testing a perfectly accurate simulated system. On any given trial, some civilians (typically 3–4) were erroneously highlighted with AiTR. While this was intended to increase visual clutter and reflect inconsistency of the AiTR, it may have decreased participants' trust and reliance on AiTR, confounding its effects on perception and decision making. Future research should consider how false alarms and trust in automation interact with perceptual and cognitive impacts of AiTR.

The present study's findings and limitations offer suggestions for future research. Current generation commercial VR HMDs offer increasingly high resolution displays and embedded eye tracking capabilities with precision and accuracy approaching laboratory-grade eye trackers (Gardony et al., 2020; Kapp et al., 2021).

Moreover, with readily accessible development platforms (e.g., Unity3D, Unreal) researchers can build rich and dynamic scenarios for the simulation, testing, and evaluation of novel and emerging AR capabilities like AiTR. Future work should continue to leverage these increasingly accessible, affordable, and capable technologies to perform psychological and user experience research. It is critically important that researchers carefully balance controlling the psychometric properties of stimuli and faithfully reflecting real-world scenarios in which to test the perceptual and cognitive impacts of factors of interest. It is also important to understand and address the technical limitations of the hardware used in the experimental design phase. In our case, we incorporated design choices like eye gaze colliders surrounding relevant scenario features that subtended a constant visual angle to address eye tracking accuracy limitations of commercial VR HMDs, designed the scenario to remain within the HMD's field-of-view, and incorporated gaze-contingent presentation of secondary stimuli to ensure visibility of peripheral stimuli. Our results also suggest that user experience metrics are important to collect alongside quantitative behavioral and gaze data. Indeed, our results demonstrated the increasing frustration with Soft Highlight AiTR was associated with increasing friend-foe discrimination error. Future research, like the present study, may observe that AR capabilities do not substantively differ when scrutinizing human performance metrics alone but do when considering survey responses alongside them.

Lastly, previous research suggests that AiTR should incorporate uncertainty visualization in tandem with target cueing (Entin et al., 1996; Cunningham et al., 2017; Larkin et al., 2019; Matzen et al., 2020). Providing analog information about the underlying algorithm's confidence embedded in or alongside the AiTR cue has been demonstrated to reduce attentional tunneling in Computer-Aided Detection tasks (Cunningham et al., 2017). However, a large body of research has demonstrated that mapping probability or confidence to different visual encoding channels to represent uncertain information can lead to biases in interpretation, especially for those lacking expertise in data visualization (Franconeri et al., 2021; Padilla et al., 2021). Thus, designers and engineers of AR AiTR systems should take care to use appropriate visual encodings that can intuitively convey uncertainty. Soft Highlight AiTR can intuitively display uncertainty, such as classification confidence or threat classification, through modifications of its fuzziness, lightness, and/or saturation (MacEachren et al., 2012). However, the present study revealed user experience decrements associated with Soft Highlight AiTR during dynamic tasks, calling into question its utility for future military AR systems. Instead, uncertainty information could be portrayed as a distinct feature adjacent to a traditional bounding box. While this may increase utilization of visual working memory (Schurgin, 2018), such visualizations can also leverage emerging best practices in uncertainty visualization, such as frequency-based visualizations (Franconeri et al., 2021; Padilla et al., 2021). Future research is needed to explore novel

AiTR designs that deliver acceptable user experience, enhanced target acquisition and decision making, and intuitive and interpretable uncertainty visualization.

Data availability statement

The datasets presented in this article are not readily available because the funding agency does not permit public release of data products. Requests to access the datasets should be directed to AG, aaron.gardony.civ@army.mil.

Ethics statement

The studies involving human participants were reviewed and approved by the United States Army Combat Capabilities Development Command Soldier Center Human Research Protection Program Office and the Tufts University Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

AG conceptualized the research and oversaw VR scenario development with critical input from GH, KR, and AS. AG and KO supervised data collection with all remaining authors contributing to data collection. AG, KO, and AK contributed to data analysis. AG and KO drafted the manuscript with feedback and revisions provided by GH, AK, and AS. All authors reviewed and approved the final manuscript.

Funding

Acknowledgments

experimental design, and Margaret Duff, Emily Weeden, and James Garijo-Garde for assistance with data collection.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Army.mil (2022). *AI-enabled ground combat vehicles demonstrate agility and synergy at PC21*. Available at: https://www.army.mil/article/251632/ai_enabled_ground_combat_vehicles_demonstrate_agility_and_synergy_at_pc21.

Åström, U. (2011). *Camouflage generator*. Available at: http://www.happyponyland.net/camogen.php?about.

Bailey, R., McNamara, A., Sudarsanam, N., and Grimm, C. (2009). Subtle gaze direction. *ACM Trans. Graph.* 28 (4), 1–14. doi:10.1145/1559755.1559757

Brunyé, T. T., and Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *Int. J. Psychophysiol.* 120, 60–68. doi:10.1016/j.ijpsycho.2017.07.008

Brunyé, T. T., and Giles, G. E. (2017). Methods for eliciting and measuring behavioral and physiological consequences of stress and uncertainty in virtual reality. *Front. Virtual Real.* (under review).

Cain, M. S., Adamo, S. H., and Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Vis. Cogn.* 21 (7), 899–921. doi:10.1080/13506285.2013.843627

Chen, J. Y. C., and Barnes, M. J. (2008). *Robotics operator performance in a military multi-tasking environment*. 3rd ACM/IEEE International Conference on Human-Robot Interaction, 279–286. doi:10.1145/1349822.1349859

Cotter, S., Lee, D., and French, A. (1999)., 76. Official Publication of the American Academy of Optometry, 631—636. doi:10.1097/00006324-199909000-00020Evaluation of a new color vision test: "color vision testing made easy*Optometry Vis. Sci.*9

Cunningham, C. A., Drew, T., and Wolfe, J. M. (2017). Analog computer-aided detection (CAD) information can be more effective than binary marks. *Atten. Percept. Psychophys.* 79 (2), 679–690. doi:10.3758/s13414-016-1250-0

Dixon, B. J., Daly, M. J., Chan, H., Vescan, A. D., Witterick, I. J., and Irish, J. C. (2013). Surgeons blinded by enhanced navigation: The effect of augmented reality on attention. *Surg. Endosc.* 27 (2), 454–461. doi:10.1007/s00464-012-2457-3

Drew, T., Cunningham, C., and Wolfe, J. M. (2012). When and why might a computer-aided detection (CAD) system interfere with visual search? An eye-tracking study. *Acad. Radiol.* 19 (10), 1260–1267. doi:10.1016/j.acra.2012.05.013

Drew, T., Guthrie, J., and Reback, I. (2020). Worse in real life: An eye-tracking examination of the cost of CAD at low prevalence. *J. Exp. Psychol. Appl.* 26 (4), 659–670. doi:10.1037/xap0000277

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Hum. Factors* 37 (1), 32–64. doi:10.1518/001872095779049543

Entin, E. B., Entin, E. E., and Serfaty, D. (1996). Optimizing aided target-recognition performance. *Proc. Hum. Factors Ergonomics Soc. Annu. Meet.* 40 (4), 233–237. doi:10.1177/154193129604000419

Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., et al. (2011). Effectiveness of computer-aided detection in community mammography practice. *JNCI J. Natl. Cancer Inst.* 103 (15), 1152–1161. doi:10.1093/jnci/djr206

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., et al. (2007). Influence of computer-aided detection on performance of screening mammography. *N. Engl. J. Med. Overseas. Ed.* 356 (14), 1399–1409. doi:10.1056/NEJMoa066099

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., and Hullman, J. (2021). The science of visual data communication: What works. *Psychol. Sci. Public Interest* 22 (3), 110–161. doi:10.1177/15291006211051956

Gardony, A. L., Lindeman, R. W., and Brunyé, T. T. (2020). "Eye-tracking for human-centered mixed reality: Promises and challenges," in *Optical architectures for displays and sensing in augmented, virtual, and mixed reality. AR, VR, mr)*. Editors B. C. Kress and C. Peroz, 11310, 230–247. SPIE. doi:10.1117/12.2542699

Geuss, M. N., Larkin, G., Swoboda, J., Yu, A., Bakdash, J., White, T., et al. (2019). *Intelligent squad weapon: Challenges to displaying and interacting with artificial intelligence in small arms weapon systems*, 11006. doi:10.1117/12.2518405

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (task Load index): Results of empirical and theoretical research,". Editors P. A. Hancock and N. Meshkati (North-Holland), 52, 139–183. doi:10.1016/S0166-4115(08)62386-9*Adv. Psychol.*

Htc (2021). *Vive - VR headsets, games, and metaverse life | United States*. Available at: https://www.vive.com/.

Kapp, S., Barz, M., Mukhametov, S., Sonntag, D., and Kuhn, J. (2021). Arett: Augmented reality eye tracking toolkit for head mounted displays. *Sensors* 21 (6), 2234. doi:10.3390/s21062234

Kim, H. K., Park, J., Choi, Y., and Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Appl. Ergon.* 69, 66–73. doi:10.1016/j.apergo.2017.12.016

Kneusel, R. T., and Mozer, M. C. (2017). Improving human-machine cooperative visual search with soft highlighting. *ACM Trans. Appl. Percept.* 15 (1)–21. doi:10.1145/3129669

Kronnect (2021). *Highlight plus | particles/effects | unity asset store*. Available at: https://assetstore.unity.com/packages/tools/particles-effects/highlight-plus-134149#publisher.

Krupinski, E. A., Nodine, C. F., and Kundel, H. L. (1993). Perceptual enhancement of tumor targets in chest X-ray images. *Percept. Psychophys.* 53 (5), 519–526. doi:10.3758/BF03205200

Larkin, G. B., Geuss, M., Yu, A. B., Rexwinkle, J., Callahan-Flintoft, C., Bakdash, J., et al. (2019). Emerging recommendations for initial automated target recognition (ATR) display considerations: Implications for concept development and evaluation. doi:10.31234/osf.io/w9qge

Lenth, R. (2020). Emmeans: Estimated marginal means, aka least-squares means. https://cran.r-project.org/package=emmeans.

Lu, W., Duh, B.-L. H., and Feiner, S. (2012). Subtle cueing for visual search in augmented reality. *IEEE international symposium on mixed and augmented reality*. New York: ACM Press, 161–166. doi:10.1109/ISMAR.2012.6402553

Lu, W., Feng, D., Feiner, S., Zhao, Q., and Duh, H. B.-L. (2014). Evaluating subtle cueing in head-worn displays. *Proc. Second Int. Symposium Chin. CHI*, 5–10. doi:10.1145/2592235.2592237

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., and Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Trans. Vis. Comput. Graph.* 18 (12), 2496–2505. doi:10.1109/TVCG.2012.279

Matzen, L. E., Stites, M. C., Howell, B. C., and Gastelum, Z. N. (2020). Effects of the accuracy and visual representation of machine learning outputs on human decision making *(No. SAND2020-11021C)*. Albuquerque, NM: Sandia National Lab.

McNamara, A., Bailey, R., and Grimm, C. (2008). Improving search task performance using subtle gaze direction. *Proc. 5th Symposium Appl. Percept. Graph. Vis.*, 51. –56. doi:10.1145/1394281.1394289

Padilla, L. M., Kay, M., and Hullman, J. (2021). "Uncertainty visualization," in *Wiley StatsRef: Statistics reference online*. Editors N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels. doi:10.1002/9781118445112.stat08296

Philpotts, L. E. (2009). Can computer-aided detection be detrimental to mammographic interpretation? *Radiology* 253 (1), 17–22. doi:10.1148/radiol.2531090689

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Ratches, J. A. (2011). Review of current aided/automatic target acquisition technology for military target acquisition tasks. *Opt. Eng.* 50 (7), 072001–072008. doi:10.1117/1.3601879

Reiner, A. J., Hollands, J. G., and Jamieson, G. A. (2017). Target detection and identification performance using an automatic target detection system. *Hum. Factors* 59 (2), 242–258. doi:10.1177/0018720816670768

Sam.gov (2022). *Artificial intelligence/machine learning software development and integration*. Available at: https://sam.gov/opp/7201d5f4370d491e8322084cf58ec4e5/view.

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Atten. Percept. Psychophys.* 80 (5), 1035–1056. doi:10.3758/s13414-018-1522-y

Singmann, H., Bolker, B., Westfall, J., Aust, F., and Ben-Shachar, M. S. (2020). *Afex: Analysis of factorial experiments*. Available at: https://cran.r-project.org/package=afex.

Sultana, F., Sufian, A., and Dutta, P. (2020). "A review of object detection models based on convolutional neural network," in *Intelligent computing: Image processing based applications*. Editors J. K. Mandal and S. Banerjee (Springer Singapore), 1–16. doi:10.1007/978-981-15-4288-6_1

Tombu, M., Ueno, K., and Lamb, M. (2016). *The effects of automatic target cueing reliability on shooting performance in a simulated military environment (DRDC Scientific Report DRDC-RDDC-2016-R036)*. Toronto, ON: Defence Research and Development Canada.

Unity Technologies (2021). *Unity real-time development platform | 3D, 2D VR & AR engine*. Available at: https://unity.com/.

US Army TRADOC (2018). *TRADOC Pamphlet 525-3-1—the U.S. Army in multi-domain operations 2028*. Available at: https://adminpubs.tradoc.army.mil/pamphlets/TP525-3-1.pdf.

Valve Corporation (2021). *Steam VR*. Available at: https://www.steamvr.com/en/.

Varjo (2021). *Most advanced virtual and mixed reality headsets for professionals—Varjo*. Varjo.Com. Available at: https://varjo.com/.

Wickens, C. D. (2005). *Tech. Rep. AHFD-05-23/NASA)*. University of Illinois, Institute of Aviation, Aviation Human Factors Division. Attentional tunneling and task management. Urbana-Champaign, IL.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Yeh, M., and Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Hum. Factors* 43 (3), 355–365. doi:10.1518/001872001775898269