From Sight to Skill: A Surgeon-Centered Augmented Reality System for Ureteroscopy Training

Jumanh Atoum[1], Fangjie Li[1], Ayberk Acar[1], Nicholas L. Kavoussi[2], and Jie Ying Wu[1]

[1] Vanderbilt University, Nashville, TN, USA
{jumanh.atoum}@vanderbilt.edu
[2] Vanderbilt University Medical Center, Nashville, TN, USA

Abstract. The number of kidney stone cases in the U.S. has tripled since 1980. Unfortunately, 23% of kidney stone removal surgeries require repeat procedures within 20 months. The high repeat surgery rate is often due to surgeons missing stones in the initial treatment. Effective training can reduce the need for re-operation, yet learning opportunities in the operating room (OR) are limited, as patient care must take priority. Augmented reality (AR) can improve the effectiveness of training in the OR by providing real-time visual feedback, but the interface must be carefully designed not to interfere with clinical workflow. Building on prior design guidelines and AR training tools, we design and evaluate the effectiveness of enhancing training through three AR gaze markers. Our AR training system tracks the expert's eye gaze and projects the marker onto the trainee's head-mounted display to provide visual guidance. Eight trainees performed a simulated ureteroscopy task of identifying kidney stones in high-fidelity kidney phantoms while guided by an expert. We record the number of stones they found, time, and eye-gaze metrics. At the end of each trial, trainees provide subjective feedback on task load and performance through the NASA-TLX questionnaire. Results show that while some gaze markers increased perceived mental demand, they enhanced engagement and performance. Gaze metrics revealed that marker shape affects cognitive load, as measured by the fixation-to-saccades ratio. By translating prior design principles into an AR-based guidance system, this work supports intraoperative training and highlights AR's potential in surgical education.

Keywords: Augmented reality · Visual guidance · Eye gaze · Kidney surgery

1 Introduction

Surgical training is a complex, time-intensive process. For some procedures, training is further limited as it occurs primarily in the operating room (OR) [10] with time and patient safety constraints [3]. In these procedures, trainees often only get limited summative verbal feedback at the end of a case. Relying mainly on verbal communication without reference information can prolong the task completion [5], and contribute to increased frustration and ineffective learning. While previous studies have explored visual aids for telehealth training [12, 19], few works have explored visual aids in intraoperative training. Previous works showed that augmented reality (AR) gaze markers increase training efficiency and improve skill acquisition in phantom procedures by providing real-time feedback [15, 22, 2]. Our previous work [2] proposed a gaze-guided training system for minimally invasive kidney stone surgery, or ureteroscopy. This work refines the AR design as a step toward using the AR training system in the OR.

The goal of ureteroscopy is to remove all kidney stones, rendering the patient "stone-free". Residual stone fragments contribute to postoperative complications and repeat surgeries, with 23% of patients requiring another surgery within 20 months [7]. Ureteroscopy training occurs primarily in real patient cases, and ineffective training contributes to high rates of residual stone fragments after surgery [16]. Our previous work demonstrated that visual guidance-aided training can improve kidney stone detection rates by up to 6.98% in phantoms [2]. We used augmented reality head-mounted displays (HMDs) to capture experts' eye gaze through HMD's built-in sensors. The tracked gaze is then projected onto the surgical trainees' HMD to provide visual guidance.

While AR has been demonstrated to improve performance in a phantom setting, poor AR design choices can increase the users' cognitive load and visual fatigue [11]. AR training systems must avoid increasing cognitive load in the already stressful OR environment. The increased cognitive load from holographic features found in [2] highlights the importance of creating and evaluating AR features through a user-centric approach. User-centered design principles have been shown to improve system usability and adoption, particularly in critical applications such as surgical training, planning, and preparation [13].

Given the importance of user-centered design, we previously conducted semi-structured interviews and co-design activities with surgical trainees to determine important factors in AR marker design [4]. In this work, we design three markers based on these findings and evaluate them in our AR-based gaze application [2]. We conducted a user study with eight trainees to evaluate gaze metrics and assessed their perceived task load through the NASA Task Load Index (NASA-TLX) questionnaires. This approach aligns with prior work [21, 17] linking eye movement features to NASA-TLX scores in surgical contexts. Our findings demonstrate that well-designed markers reduce cognitive load while enhancing task engagement, underscoring the potential of AR-driven solutions to optimize surgical education as well as incorporating user feedback into the design process. By integrating user-centered design-based recommendations for AR-based features, we amplify the benefits of integrating AR-HMD in intraoperative training.

2 Methods

We summarize the design guidelines to provide context on the development of the gaze markers: (1) track and project the expert surgeons' eye gaze, (2) remain static between significant gaze shifts to minimize distraction, (3) have contrastive colors with the body tissue, (4) be appropriately sized for the surgical application, and (5) not block critical anatomy.

Development of Novel Markers for Gaze Tracking Application Based on the guidelines above, we iterated with a board-certified urologist who specializes in kidney stone surgery to refine the three AR markers. We created the markers using Blender [9] and then deployed them to Unity (2021.3.4.f1) (Unity Technologies, San Francisco, CA). Below are our main design considerations:

Shape We include a baseline simple dot marker (M1) (Fig. 1a) to have a comparable marker to previous studies [22, 2]. From our co-designs, we see a preference for shapes that do not obscure anatomy and are easily recognizable by surgical trainees. Thus, we designed the second marker (M2) (Fig. 1b) to highlight a region without obscuring the center. In M2, the small dot represents the actual gaze point, and the circle is intended to direct the trainee's attention to the focus area. We created the third marker (M3) (Fig. 1c) to test the trainee's feedback that seeing the direction of the expert's eye gaze movement could be helpful. The largest circle represents the expert's gaze at the current time step.

Color The previous study showed that green/blue are the colors with the highest contrast against tissue color. For our marker color, we chose a bright blue color that has high contrast with light pink and red, the color of our phantoms, which to some extent matches the kidney's collecting system. We chose blue over green to avoid common red/green color blindness.

Size The design guidelines did not specify the size of AR markers. However, results from the previous study highlighted that markers should not be too large to obscure anatomy nor too small to follow [4]. To balance visibility and obstruction, M1 had a 0.5 cm radius, M2 had a 1.0 cm radius, and M3 fit within a 2.0 cm $\times$ 2.0 cm box. This design choice aligns with the 1.5 cm - 2.0 cm eye gaze error range of HoloLens.

(a) Dot marker (M1) (b) Target marker (M2) (c) Trail marker (M3)

Fig. 1: The three devised AR markers informed by the guidelines generated from our previous semi-structured user study with co-design activity. The marker sizes are not reflected in the figure.

Eye Gaze Tracking based AR Application We developed a real-time AR application to share expert surgeons' eye gaze with trainees. The application was built in Unity using MRTK2 and Photon Unity Networking (PUN). The application captures gaze data from HoloLens 2's onboard sensors. Before starting the study, the user performs the HoloLens 2 eye calibration to ensure accurate gaze tracking. The study starts with both users standing side by side, facing a surgical monitor localized by four ArUco markers, as seen in Fig. 2. The centers of the bottom and left-side markers are used to calculate the monitor's width and height, respectively, as shown by the surrounding boundary in Fig. 2. The monitor's virtual position is set by averaging the markers' locations, allowing each HoloLens to project gaze data onto a matched virtual screen. The expert's gaze (${}^{exp}T_{gze}$) and expert frame to screen frame transform (${}^{exp}T_{scr}$) are sent via PUN; the trainee's HoloLens computes the expert's gaze on the virtual screen using the transform ${}^{trn}T_{scr} \times {}^{exp}T_{scr}^{-1} \times {}^{exp}T_{gze}$. This setup ensures visibility from all angles and minimizes marker bias. A computer-based remote-control tool manages marker display during studies.

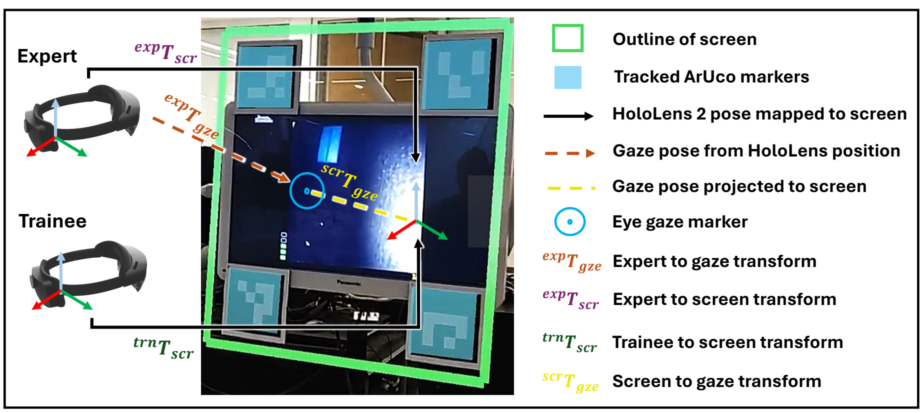


Fig. 2: Diagram showing the AR application's eye gaze overlay process with the virtual monitor fitting and defining the relevant frame transformations.

3 Experimental Setup

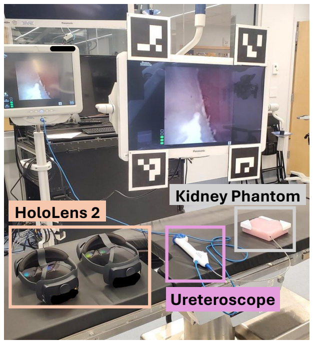
Participant Recruitment and Demographics We recruited eight surgical trainees, all from the urology department at Vanderbilt University Medical Center. Four participants identified as female and four as male. The participants' expertise ranged from one to four post-graduate years of practice (PGY). In total, two participants were PGY1, one was PGY3, and five were PGY4. Our study received approval from the hospital's Institutional Review Board (IRB 231997). Participant recruitment ran from October 2024 to February 2025.

High Fidelity Kidney Phantom For this study, we prepared four distinct kidney phantoms by casting silicone using 3D-printed kidney collecting systems from patient CT scans, and an outer mold (Fig. 3a). More details of the phantom creation process can be found in [1]. An expert surgeon validated the complexity and exploration difficulty of the renal collecting systems in each phantom as similar to each other. The experimental setup with all phantoms and the OR environment is shown in Fig. 3b.

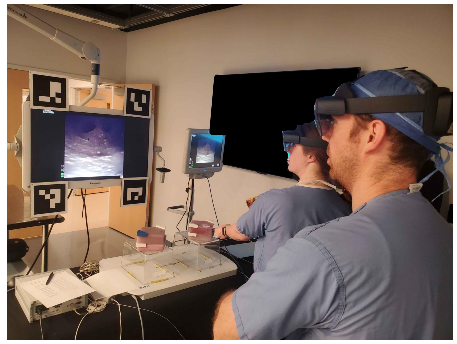
User Study Protocol To evaluate our marker designs, we conducted a user study with an expert surgeon and eight surgical trainees. In each trial, a surgical trainee was asked to explore a kidney phantom using a LithoVue ureteroscope (Boston Scientific, Marlborough, MA) with guidance from the expert surgeon. The trainee was asked to fully explore the kidney and report the number of stones they found. They were not told that each phantom contained five stones. Each surgical trainee performed four trials — once with each of the three marker designs and once with no marker. In the no marker cases, only verbal guidance was provided, as is the current standard of care. In the marker cases, both verbal and visual guidance were provided. The trainees wore the HoloLens 2 in all tasks. For each user, the order of visual guidance markers and phantoms was randomized to avoid learning bias.

The user study started with both users wearing HoloLenses and performing built-in eye gaze calibration. We validated the expert's eye gaze projection by having the expert look at a grid while the surgical trainee identified which box in the grid the expert was looking at, similar to [2]. The gaze projection error was 1.5 cm. Then, we waited for a cue from the surgical trainee to start the task, at which point a timer began. We defined task completion as the moment when surgical trainees decided that they were done exploring the phantom. In each task, the surgical trainees navigated the ureteroscope with expert guidance and identified stones verbally, confirmed by both the expert and the in-room researcher. We recorded the number of stones found and the time taken to complete the task. After completing each task, surgical trainees filled out a NASA-TLX questionnaire [14]. The form assessed six categories: mental demand, physical demand, temporal demand, performance, effort, and frustration, each rated on a scale from 1 (best) to 20 (worst). We measured task completion time, the number of stones found, survey responses, and gaze metrics for each user and task. We also asked the trainees which marker they preferred.

For gaze analysis, we used the metrics with the most significant statistical results in [18], namely, total distance traveled by eye gaze, fixation-to-saccades ratio, and gaze area in the area of interest (AOI). The total distance traveled by eye gaze measures all movement on the virtual plane. We used the same definition of fixation, which is limited eye movement for at least 300 ms within a circle of 1.5 cm diameter, and overall fixation time is the sum of time spent in each fixation period, while the remainder time is considered saccades. Finally, we set the AOI to be part of the surgical monitor displaying the endoscopic video stream, with a size of 35.5 cm $\times$ 35.5 cm.



(a) Experimental setup



(b) User study setup

Fig. 3: (a) Our user study setup includes a ureteroscope for kidney phantom navigation and ArUco markers for monitor registration. (b) Expert and trainee performing simulated ureteroscopy.

4 Results

More than 62% of participants (n = 5) chose M2 as their favorite marker, (n = 2, 25%) chose M1, and (n = 1, 13%) chose M3. No participant chose verbal guidance alone as their preferred setup. One trainee noted, "[M2] is very subtle; I don't need to focus on it the whole time, but it still provides the needed guidance." Another trainee mentioned that the marker size can obstruct their view, "[M3] is too large compared to [M1] and [M2], making it difficult to focus on what is behind it." Another trainee mentioned that "It was difficult to locate [M1] some times during the task." All trainees agreed that the color of the marker is contrastive, making the marker visible at all times.

Time and Stone Count From Table 1, we observe that M1 resulted in the highest number of stones found and a shorter completion time compared to the no marker condition. Similarly, M2 achieved a comparable number of stones found to M1 but had a longer average completion time than the no marker condition. In contrast, M3 had the lowest number of stones found and the highest completion time compared to M1 and M2. These performance differences are attributed to marker effectiveness rather than expertise, since all participants received all three levels of guidance, and each trainee's performance is compared to their own verbal guidance baseline. Additionally, the difficulty level of the phantoms was consistent, despite some anatomical variability. These phantoms reflect typical anatomy, and our analysis of users' performance averaged over each phantom showed no significant differences.

Gaze Metrics Table 1 shows that adding a marker increased the *fixation-to-saccade ratio*, indicating higher cognitive load. Among the three markers, M2 caused the smallest increase. No marker condition had the lowest cognitive load. We observe that M2 had the lowest *total distance* compared to M1, M3, and no marker conditions. We also observed that M3 had the smallest gaze area in AOI.

A two-tailed Welch's t-test [23] showed no significant difference between no marker and each of the three markers or among the three markers themselves, except for the fixation-to-saccade ratio, which differed significantly between M2 and M3 ($p = 0.07$).

Table 1: Mean (standard deviation) of gaze and objective metrics for surgical trainees under the three markers.

	No Marker	M1	M2	M3
Number of Stones ↑	4.4 (0.7)	**4.9** (0.4)	4.6 (0.5)	4.4 (0.5)
Completion Time (s) ↓	73.7 (25.5)	**72.1** (21.9)	84.7 (23.2)	88.6 (25.5)
Total Distance (cm) ↓	523.4 (173.1)	590.1 (354.7)	**502.0** (184.0)	531.3 (217.4)
$\frac{Fixation}{Saccades}$ ↓	**40.7** (54.7)	51.3 (49.5)	42.1 (61.7)	52.7 (36.0)
Area in AOI (cm^2) ↓	132.2 (35.4)	148.2 (58.7)	138.9 (37.0)	**122.3** (48.6)

Task Engagement based on Users' Self-Evaluation We observe improvement on all metrics over no marker across different marker shapes, except for *physical demand*. We hypothesize that visual guidance increased physical demand by requiring additional hand-eye coordination for multi-tasking. This multi-tasking is attributed to surgical trainees processing visual cues while simultaneously controlling the ureteroscope.

In Table 2, M1 had the highest metric scores except for *performance* and *physical demand*, while M2 had the highest *performance* score. Additionally, M2 had lower *effort* and *frustration* compared to no marker. M3 had a lower *frustration* score than M2, but had a higher *effort* score. In the case of *mental demand*, M2 had a higher score than no marker but a lower score than M3.

Table 2: Mean (standard deviation) NASA-TLX survey results. Lower is better.

	No Marker	**M1**	**M2**	**M3**
Mental Demand ↓	8.9 (4.0)	**7.1** (2.1)	9.5 (3.2)	10.7 (5.2)
Physical Demand ↓	**6.7** (2.8)	6.9 (3.0)	7.6 (4.3)	7.4 (4.4)
Temporal Demand ↓	10.3 (3.6)	**8.6** (3.9)	10.3 (4.6)	9.0 (4.4)
Performance ↓	8.1 (2.3)	6.3 (3.3)	**6.0** (3.3)	8.4 (4.6)
Effort ↓	11.3 (3.1)	**7.7** (2.7)	10.3 (4.0)	10.6 (5.2)
Frustration ↓	9.7 (3.3)	**6.1** (2.3)	9.0 (4.5)	8.7 (4.9)

5 Discussion and Conclusion

We assess the perceived effects of all three markers by jointly analyzing subjective and objective metrics.

M1 was associated with the highest fixation-to-saccade ratio, suggesting relatively higher cognitive load compared to other markers. This can be attributed to the effort required to follow a smaller marker. From Table 2, the combination of lower performance, effort, and frustration can indicate a lower task engagement [8]. In terms of gaze behavior, M1 resulted in broader but less targeted exploration of the kidney.

M2 showed the lowest total gaze distance and the second lowest gaze area in AOI, suggesting that users explored a larger portion of the kidney while maintaining an efficient gaze path. Efficient gaze path is an indication of higher expertise. In terms of users' self-evaluation, M2 is associated with increased engagement and deeper learning [8], as it had the best performance score and low to moderate effort and frustration scores. Frustration, in particular, has been linked to discrepancies between skill and task demands [20] and can support long-term learning when appropriately regulated [6]. Although M2 had a higher mental demand score than M1, it had a lower fixation-to-saccade ratio. This suggests that the task remained appropriately challenging while maintaining a relatively low cognitive load. M2 did not alter the temporal demand compared to M1 and M3. This suggests sustained task engagement without urgency.

M3 showed the smallest gaze area in AOI compared to all markers, which may reflect focused gaze behavior; however, this outcome might result from users attending more to the marker's complex visual composition, rather than its intended referent. M3's visual design included multiple circles, potentially diverting the trainees' attention. Although users had a limited gaze spread while using M3, total gaze distance remained relatively high. In terms of self-evaluation metrics, M3 had the highest mental demand score, with high frustration and effort, suggesting a comparatively higher cognitive burden during task execution. In contrast to M2 and M1, M3 did not show clear benefits across objective or subjective metrics but did distinctly influence attention patterns.

Our results support the user's preference of M2. Users' NASA-TLX indicated that they thought they had the best performance, which is supported by their more expert-like gaze metrics. Although users did not find the most stones with M2, expert-like gaze behavior could indicate better skill acquisition.

Our study is limited to kidney ureteroscopy as the primary task for assessing cognitive load and task engagement. Future work should explore navigation-based endoscopic procedures and evaluate the effectiveness of visual markers across different surgical tasks. Additionally, our user study was conducted on a single type of phantom model, restricting the generalizability of the findings to other endoscopic procedures. Expanding future studies to include a wider variety of phantom models, such as those used in laparoscopic renal surgery, would enhance the applicability of our results.

Another limitation is the relatively small sample size of participants, which reduces the statistical power and limits the generalizability of our findings across broader populations. Future research should include a larger and more diverse participant pool to improve the robustness of the conclusions. We also aim to investigate the impact of the preferred gaze marker on surgical trainees' skill retention rate and compare active learning levels with verbal guidance. Addressing these challenges will contribute to developing more effective training models, ultimately enhancing surgical education and improving clinical outcomes.

In conclusion, this work evaluated the impact of three gaze markers in an AR gaze-based training application. Each trainee performed ureteroscopy on four phantoms, using each of the gaze markers with verbal feedback or verbal feedback alone. We report trainees' task performance, gaze metrics, and self-reported task load (NASA-TLX). Our results show that well-designed markers reduce cognitive load and enhance task engagement. By emphasizing surgeon-centered AR design, our work establishes a foundation for AR-assisted surgical training and medical education.

Acknowledgment This study was partially supported by the NIBIB of the NIH Grant 1R21EB035783 and the Wellcome Leap SAVE Program.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acar, A., Atoum, J., Connor, P.S., Pierre, C., Lynch, C.N., Kavoussi, N.L., Wu, J.Y.: NAVIUS: Navigated Augmented Reality Visualization for Ureteroscopic Surgery. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025 (2025)
2. Acar, A., Atoum, J., Reed, A., Li, Y., Kavoussi, N., Wu, J.Y.: Intraoperative gaze guidance with mixed reality. Healthcare technology letters. (2023-12-13)
3. Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., Kneebone, R.: The impact of stress on surgical performance: a systematic review of the literature. Surgery **147**(3), 318–330 (2010)
4. Atoum, J., Park, J., Akter, M., Kavoussi, N., Wisniewski, P., Wu, J.Y.: Focus on the experts: Co-designing an augmented reality eye-gaze tracking system with surgical trainees to improve endoscopic instruction. arXiv preprint arXiv:2506.21896 (2025), https://arxiv.org/abs/2506.21896
5. Bai, H., Sasikumar, P., Yang, J., Billinghurst, M.: A user study on mixed reality remote collaboration with eye gaze and hand gesture sharing. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–13 (2020)
6. Bjork, R.A., Bjork, E.L., et al.: A new theory of disuse and an old theory of stimulus fluctuation. From learning processes to cognitive processes: Essays in honor of William K. Estes **2**, 35–67 (1992)
7. Brain, E., Geraghty, R.M., Lovegrove, C.E., Yang, B., Somani, B.K.: Natural history of post-treatment kidney stone fragments: A systematic review and meta-analysis. The Journal of Urology **206**(3), 526–538 (2021)
8. Chi, M.T., Wylie, R.: The icap framework: Linking cognitive engagement to active learning outcomes. Educational psychologist **49**(4), 219–243 (2014)
9. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org
10. Drew, P., Cule, N., Gough, M., Heer, K., Monson, J., Lee, P., Kerin, M., Duthie, G.: Optimal education techniques for basic surgical trainees: lessons from education theory. Journal of the Royal College of Surgeons of Edinburgh **44**(1), 55–56 (1999)

11. Elias, Z.M., Batumalai, U.M., Azmi, A.N.H.: Virtual reality games on accommodation and convergence. Applied ergonomics **81**, 102879 (2019)
12. Feng, Y., Li, K., Semsar, A., McGowan, H., Mun, J., Zahiri, H.R., George, I., Park, A., Kleinsmith, A., Mentis, H.M.: Communication cost of single-user gesturing tool in laparoscopic surgical training. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2019)
13. Guan, Z., Lee, S., Cuddihy, E., Ramey, J.: The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 1253–1262. CHI '06, Association for Computing Machinery, New York, NY, USA (2006)
14. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier (1988)
15. Kim, D., Choi, S., Choi, J., Shin, H., Sohn, K.: Visual fatigue monitoring system based on eye-movement and eye-blink detection. In: Stereoscopic Displays and Applications XXII. vol. 7863, pp. 35–42. SPIE (2011)
16. Kim, H.J., Daignault-Newton, S., DiBianco, J.M., Conrado, B., Jafri, S.M., Seifman, B., Konheim, J., Dauw, C.A., Ghani, K.R., Collaborative, M.U.S.I.: Real-world practice stone-free rates after ureteroscopy: variation and outcomes in a surgical collaborative. European urology focus **9**(5), 773–780 (2023)
17. Larki, A.A., Shojaei, A., Delrobaei, M.: Towards cognitive load assessment using electrooculography measures. arXiv preprint arXiv:2312.11418 (2023)
18. Li, Y., Reed, A., Kavoussi, N., Wu, J.Y.: Eye gaze metrics for skill assessment and feedback in kidney stone surgery. International Journal of Computer Assisted Radiology and Surgery pp. 1–8 (2023)
19. Maria, S., Mentis, H.M., Canlorbe, G., Avellino, I.: Supporting collaborative discussions in surgical teleconsulting through augmented reality head mounted displays. In: Proceedings of the 2023 CHI conference on human factors in computing systems. pp. 1–13 (2023)
20. Marsh, E.J., Butler, A.C.: Memory in educational settings. In: Reisberg, D. (ed.) Oxford Handbook of Cognitive Psychology, pp. 299–311. Oxford University Press (2014)
21. Mazur, L.M., Mosaly, P.R., Hoyle, L.M., Jones, E.L., Marks, L.B.: Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. Practical radiation oncology **3**(4), e171–e177 (2013)
22. Melnyk, R., Campbell, T., Holler, T., Cameron, K., Saba, P., Witthaus, M.W., Joseph, J., Ghazi, A.: See like an Expert: Gaze-Augmented Training Enhances Skill Acquisition in a Virtual Reality Robotic Suturing Task. Journal of Endourology **35**(3), 376–382 (3 2021). https://doi.org/10.1089/end.2020.0445
23. Welch, B.L.: The significance of the difference between two means when the population variances are unequal. Biometrika **29**(3/4), 350–362 (1938)