Mixed Reality Communication for Medical Procedures: Teaching the Placement of a Central Venous Catheter

Manuel Rebol[*]
American University, Graz University of Technology

Krzysztof Pietroszek[†]
American University

Claudia Ranniger[‡]
George Washington University

Colton Hood[§]
George Washington University

Adam Rutenberg[¶]
George Washington University

Neal Sikka[‖]
George Washington University

David Li[**]
George Washington University

Christian Gütl[††]
Graz University of Technology

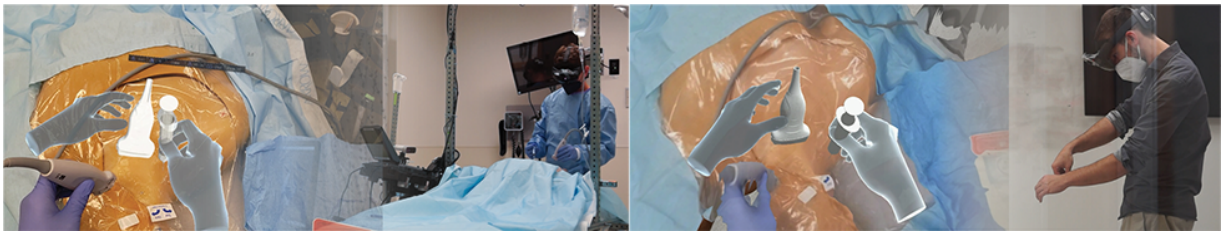


Figure 1: Mixed reality communication for medical procedures. We present a mixed reality communication system. A remote expert (right) guides a local operator (left) through the placement of a central venous catheter using augmented objects.

ABSTRACT

Medical procedures are an essential part of healthcare delivery, and the acquisition of procedural skills is a critical component of medical education. Unfortunately, procedural skill is not evenly distributed among medical providers. Skills may vary within departments or institutions, and across geographic regions, depending on the provider's training and ongoing experience. We present a mixed reality real-time communication system to increase access to procedural skill training and to improve remote emergency assistance. Our system allows a remote expert to guide a local operator through a medical procedure. RGBD cameras capture a volumetric view of the local scene including the patient, the operator, and the medical equipment. The volumetric capture is augmented onto the remote expert's view to allow the expert to spatially guide the local operator using visual and verbal instructions. We evaluated our mixed reality communication system in a study in which experts teach the ultrasound-guided placement of a central venous catheter (CVC) to students in a simulation setting. The study compares state-of-the-art video communication against our system. The results indicate that our system enhances and offers new possibilities for visual communication compared to video teleconference-based training.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Computing methodologies—Computer vision—Computer vision problems—Reconstruction Social and professional topics—Medical information policy—Medical technologies—Remote medicine; Telehealth; Volumetric communication;

[*]e-mail: mrebol@american.edu
[†]e-mail: pietrosz@american.edu
[‡]e-mail: ranniger@gwu.edu
[§]e-mail: chood@mfa.gwu.edu
[¶]e-mail: arutenberg@mfa.gwu.edu
[‖]e-mail: nsikka@mfa.gwu.edu
[**]e-mail: dli@mfa.gwu.edu
[††]e-mail: c.guetl@tugraz.at

1 INTRODUCTION

Patient care and procedural skills together form one of the Accreditation Council for Graduate Medical Education's (ACGME's) six core competencies for a practicing physician [13]. Learning procedural skills typically requires that a trainee and an experienced medical professional are co-located. Training must be repeated periodically if the trainee does not perform the procedure regularly during training. Furthermore, the skills required to perform the procedure should be practiced periodically after training to avoid degradation [17], especially if the operator does not perform the procedure regularly in his or her daily medical practice.

Multiple educational frameworks describe the acquisition of new procedural skills, but all have in common the iterative development of skill in a less experienced operator under the supervision and evaluation of a more experienced operator. Unfortunately, adequate procedural skill training is not always available to all medical providers. Thus, medical providers' skills may vary depending on the department, institution, and geographic region. Nevertheless, some providers with limited experience may be placed in an emergency situation in which a procedure must be performed immediately [25]. Consequently, there is an ongoing need to improve procedural skill acquisition and to provide remote assistance for medical providers with limited prior procedural experience or who work in resource-poor settings. Additionally, due to the geographical distribution of medical personnel, it is often difficult to arrange co-located training once the medical practitioner has completed initial medical training, especially for those who work in a remote areas, *e.g.* in a critical access hospital.

In order to address this issue, we will describe the design and implementation of a prototype real-time mixed-reality volumetric communication system that supports the acquisition of procedural skills for remote medical trainees. The system allows a remote expert to train and assist a medical trainee in learning a medical procedure without a need to be co-located with the trainee. We show examples of the different views of our communication system in

2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)
https://doi.org/10.1109/ISMAR55827.2022.00050

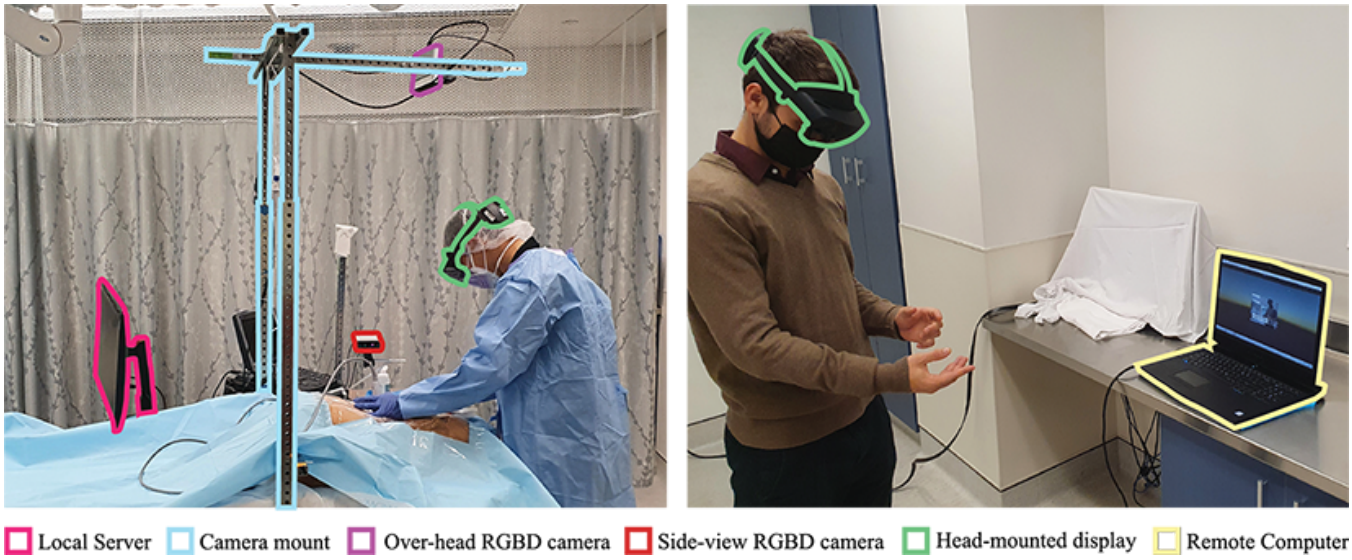


Figure 2: MR system overview for CVC placement. The local operator (left image) places a CVC while being assisted from a remote expert (right image) using the real-time mixed-reality communication system. We highlight the hardware components of the system.

Figure 1. We use the life-saving ultra-sound guided central venous catheter (US-CVC) placement procedure as an example procedure for our system design. We compare our mixed reality communication system against traditional video assistance in a user study. The participants complete the NASA Task Load Index (NASA-TLX) [11] and open-ended surveys.

2 Related Work

Mixed reality (MR) poses a significant opportunity to enhance simulation-based training (SBT). Si et al. found that AR-based training simulations were able to accurately represent neurosurgical procedures, which is essential for a novice's comprehension and application of such training [32]. Shenai et. al. used the Virtual interactive presence and augmented reality (VIPAR) tool to provide telepresence virtual expert assistance during neurosurgical procedures. Using stereoscopic microscopes, both surgeon and expert were able to see both the surgical field and each other's hands, with the remote expert able to provide visual and verbal guidance [31]. A simplified, non-stereoscopic system was subsequently used to support remote pediatric neurosurgical procedures with success [6]. Similarly, Rojas-Muñoz et al. found that medical students were able to make incisions with greater accuracy using a telepresence AR system [29].

A remote surgical assistance MR/AR system, known as ARTEMIS, was recently developed by Gasques et al. [7]. This system provides a 3D representation of the expert to the student and is able to overcome many of the communication issues inherent with remote SBTs. For example, a remote expert is able to provide live annotations of the surgical field, while providing 3D hand gestures that can be visualized by the trainee and ultimately assist with the procedure. Initial evaluation of the system consisted primarily of qualitative feedback from study participants. The researchers found that novice trainees were able to successfully complete several complex surgical procedures while using the system. It is difficult, however, to assess how this system affected a trainee's cognitive load given the qualitative nature of the study. We propose a more affordable and less hardware-complex mixed reality communication solution compared to Gasques et al. Moreover, remote ultrasound (US) training adds additional challenges to the communication system [14, 24, 33]. Mahmood *et al*. [18] proposed how US views can be used effectively in AR.

A first-person view of the procedural space has notable value in communicating elements of a trainee's environment to a remote expert. Some AR and MR systems have integrated this feature, which has been found useful by remote experts instructing trainees [7]. Hand gestures themselves provide important non-verbal cues that may also add greater value to the trainee and remote expert interaction. Gestures may be used to direct trainees to a specific area of interest within the procedural space, or how to manipulate tools relevant to the procedure. Few studies have aimed to categorize such gestures in the context of surgical maneuvers for SBTs that involve MR or AR systems. Gesture recognition may increase the complexity of such a system but has utility in remote collaborative environments [36]. Complex gestures, such as those found in medical procedure training, may be further analyzed into subcategories to add context during a remote collaboration.

Ultimately, the preceding studies illustrate some current topics in the SBT MR and AR literature. Despite these novel findings, it is uncommon for studies to provide a synthesized and comprehensive solution that not only builds on the current state of AR/MR but also employs validated instrumental tools like the NASA-TLX. Also, studies rarely employ standardized gesture analysis within the context of medical SBT MR/AR systems. By employing validated tools, MR/AR systems may be rigorously tested for viability within medical SBT environments and reach the threshold for influencing patient care.

To allow for volumetric communication, 3D scene reconstruction algorithms are used to combine multiple RGBD views to create a volumetric mesh. Most recent algorithms utilize machine learning on large data sets [4, 10] for high quality reconstruction of RGBD camera [2, 3, 16, 35] footage. Yet, high-quality reconstruction algorithms are too slow [37] for real-time communication. To overcome this issue, real-time 3D surface reconstruction has been proposed [8,15,38]. Chen *et al*. [5] achieve 24 Hz, for static scenes. For dynamic scenes, Yu *et al*. [39] propose a real-time volumetric reconstruction algorithm to capture humans. Meertis *et al*. [19] capture scenes with grid-based spatial and temporal depth map lookups for live volumetric view generation using desktop computers with high-end GPUs. Meertis *et al*. filter the RGBD point cloud using a moving least squares (MLS) algorithm. Yet, the maximum time to create a mesh takes 167ms. We identified the need for low-latency and low-bandwidth visual communication that runs on mobile devices

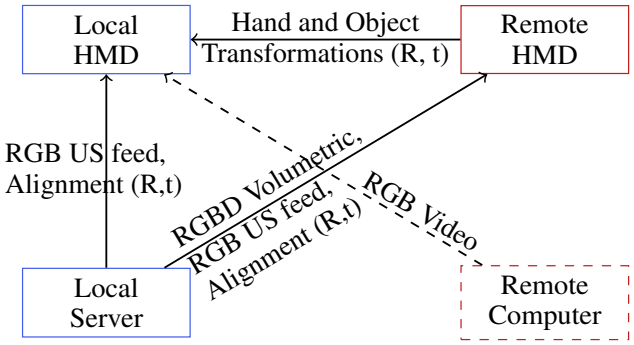


Figure 3: Device and data exchange overview. We illustrate the four main communication devices using boxes. The arrows indicate the information transfer between the devices. Local and remote devices are highlighted in blue and red, respectively. The optional Remote Computer provides a video feed of the remote scene.

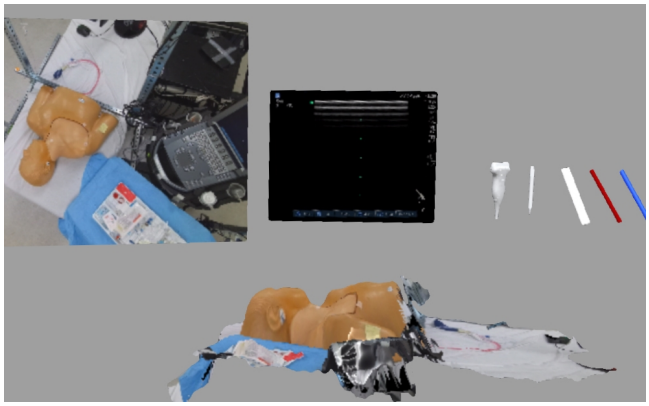


Figure 4: Remote view. The remote virtual workspace consists of the volumetric view (bottom center), the video feed (left), the ultrasound feed (top center), and the virtual objects (right). The virtual workspace can be placed in any room.

for medical training and emergencies. Our proposed mesh generation algorithm has less than 1/30s latency and runs on mobile GPUs that are found in state-of-the-art head-mounted displays (HMDs). Moreover, our approach supports low-bandwidth connections.

3 MR System Design

We propose a MR communication system for US-CVC training and emergency assistance to tackle the problem of unequal distribution of healthcare providers and procedural skills. We identified the initial design requirements for the mixed reality (MR) communication system in an elicitation study. In the elicitation study we analyzed in-person US-CVC training. We identified the need for spatial information, voice & hands-free communication, aligned hand tracking, and virtual objects. Then, we started implementing and iteratively received feedback from medical experts to improve the system including the user interface and the augmented workspace setup.

During our system design phase, we tested different means of visual communication including a drawing and a pointing feature. However, we found virtual objects to be more useful for the US-CVC procedure. We also experimented with different alignment methods including markers and HTC Vive trackers. We found that our point correspondence method provides the best tradeoff between accuracy and setup time.

We designed our system as a two-way real-time volumetric-based telepresence. It was designed to teach and support the US-CVC procedure, which is a procedure to place a large catheter into the central venous circulation. The procedure requires external landmark identification and psycho-motor skills to combine hand movements with information provided by ultrasound. Ultrasound is used to identify relevant anatomy and provides image guidance to facilitate a needle in puncturing the appropriate blood vessel which prevents injury. The two parties that our system is designed for are the remote expert and the local operator of the US-CVC procedure. The system supports a one-to-one connection between remote expert and the local operator.

3.1 System Components

We address the requirements for mixed reality guidance of a US-CVC by designing the following components:

- We use a head-mounted augmented reality display for presenting the remote guidance to the local operator. The advantage of this technology is that information from the remote operator can directly be augmented on the local operator's view. Thus, the local operator can focus on the procedure and does not need to use hands to operate the technology.
- Similarly, the remote instructor wears an AR head-mounted display (HMD) to be able to view a volumetric representation of the scene. Furthermore, the gestures and the interaction with the scene are captured by the HMD such that natural on-scene-like interaction is supported.
- We deploy two volumetric cameras at the local site to present the local scene to the remote operator. We render the captured volumetric scene on the HMD the remote operator is wearing.
- We use a microphone and speakers for voice communication between the remote expert and the local operator.

We illustrate the system including actors, devices, and communication flows in Figure 2 and Figure 3.

3.2 Devices

We use the Microsoft Hololens 2 HMD for augmenting the view of the local operator and for presenting the local view to the remote expert. In addition, the Hololens 2 captures the remote instructor's gestures that are sent to the local operator. The local scene is recorded by the volumetric camera Microsoft Azure Kinect. The procedure-specific ultrasound (US) feed is provided by a Sonosite M-Turbo machine. We deploy a small form factor computer on the local site to process the RGBD capture of the Azure Kinect camera and the US feed in real-time. Moreover, the computer acts as server that forwards the data between the Hololenses and the remote computer.

3.3 Views

The physical environment, as well as augmented information, are visible through the visor of the Hololens for the local operator and the remote expert. The augmented information allows for visual communication and interaction between the two actors.

Remote View The remotely located instructor receives the volumetric view recorded by cameras at the local site. It is augmented in 3D space on the instructor's HMD. The volumetric view is the center of the augmented workspace. It is placed in the middle of the room the remote expert is situated in. The room needs to be empty because the remote expert needs space to interact with the volumetric view. We present the remote view in Figure 4.

Besides the volumetric view, the remote expert also sees a 2D video feed. The 2D video feed is captured by the same camera that also captures the volumetric view. Moreover, the remote expert sees the ultrasound feed from the local operators ultra-sound machine. The remote expert is able to manipulate a small set of virtual objects that can be used to instruct the local operator. These include abstract objects such as cuboids and cylinders as well as realistic renderings of medical tools used for the US-CVC procedure. These may be selected by the instructor, manipulated in space, resized, and shown to the local operator.

On top of the remote expert's hands, an augmented hand model is shown. This is the same hand model that is also shown to the local operator. It gives the remote expert a better understanding of how hand gestures look augmented on the local view. When the remote experts look at their palm, a hand menu is shown which gives them the options to switch between the cameras, enable and disable the virtual objects, and switch to long reach virtual arms.

Local View The physical environment plays an important role in the local view. The local operator needs to focus mostly on the physical workspace including the patient and medical instruments. Thus, the augmented visuals should only contain the most important information needed to get the required assistance from the remote expert. Moreover, we managed to eliminate all device interaction with the MR interface for the local operator. We present the local view in Figure 5.

The augmented information for the local operator includes two feeds in a static position. A video feed showing the remote expert and the ultrasound feed. Both feeds are located right above the physical area of the procedure. The reason for the augmented US feed in our system is to provide the local operator with information close to the procedural area. This allows the local operator to make smaller changes in focus when alternating attention between the US and physical area of operation while coordinating hand movements. Alternatively, the local operator can use the physical US display. The position of the US feed was determined after consultation with US-CVC experts. In addition to the static feeds, the local operator also sees the virtual objects and the instructors virtual hand [23] model augmented by their HMD. The virtual objects can be moved by the remote expert. The virtual hand model moves as the remote operators hands move.

3.4 Interaction

The visual interaction happens in both communication directions. The local operator's body language is captured through the Azure Kinect cameras and presented as a volumetric view to the remote expert. The remote expert has two options to guide the local operator:

- The expert can use hand gestures [26, 27], which are captured by the Hololens 2 camera system.
- The expert can use virtual objects, which can be manipulated by grabbing them with hands.

The hand model can be used to give directions in the form of *e.g.* pointing, showing how to hold instruments, and showing angles. The virtual objects can be used to show how to use them correctly including how to hold them and where to place them on the patient, and how to manipulate them in space.

4 MR System Implementation

The three essential parts of our MR system are the network communication between the nodes, the 3D scene reconstruction and the alignment of the views between the two actors. For network connection, we use a peer-to-peer as well as a client-server architecture depending on the data communication types. The 3D scene captured by two Azure Kinect RGBD cameras is reconstructed using a grid mesh topology. The views are aligned between the remote and local operators using 3D point correspondences and head tracking to enable volumetric visual communication.

Figure 5: Local view. The physical workspace of the local operator includes the patient, the US device, and CVC-specific objects. The virtual view (highlighted using green borders for illustration) is used by the local operator to receive visual guidance from a remote expert.

4.1 Network Setup and Data Communication Between Devices

From a networks perspective, we have four nodes in our system. We have the local computer, the local Hololens, the remote computer, and the remote Hololens. Each of the four nodes runs software we developed in Unity. The local computer also acts as server for initiating the peer-to-peer connections and establishing WebSocket connections. We show a diagram of the network and the data flow in Figure 3.

From a network perspective, we implement the Mixed Reality WebRTC client to manage the video data transfer from local server to remote Hololens, from local server to local Hololens and from remote computer to local Hololens. Moreover, we establish WebSocket connections between the local server, the local Hololens, and the remote Hololens. Through the WebSocket connection, we forward depth, transformation, and alignment data. We explain the data communication between the four devices below.

Local Server The local server provides three main services: network connection handling, local view capture, and local Hololens rending. As a network server, the local server initiates the WebRTC peer-to-peer connection and handles all WebSocket connections. We support unicast and multicast delivery of WebSocket data. The local view capture consists of the ultrasound feed and the volumetric view. Moreover, the local server renders the view for the local Hololens and sends it over network to the device.

The volumetric view, consisting of color and depth frames are sent over the network through WebRTC and WebSocket channels, respectively. The frame number is encoded in both streams for synchronization purposes. We synchronize the color and the depth frames once received them from the Kinect camera and at the HMD of the remote operator. The mesh consistency is enforced before it is sent to the remote Hololens. The ultrasound input is transmitted as a video feed to the local and the remote Hololens. In addition to volumetric and US data, alignment information between the two Azure Kinect cameras is sent to the local and remote Hololens.

In terms of hardware, the local server is equipped with a state-of-the-art consumer CPU and GPU to allow for real-time processing. Two Microsoft Azure Kinect RGBD cameras and the ultrasound machine are connected. The cameras capture color and the depth frames of the scene at 30 FPS.

Remote HMD The remote instructor receives color and depth frames from the local cameras and the US feed. The transformation between the two Azure Kinect cameras is sent to render the volumetric view at the remote site. Depth and color frames are received at 15 FPS. We found the following buffer and synchronization strategy to work best for our system. The remote HMD waits for up to 100 ms before rendering, if color and depth frames arrive out-of-order. After 100 ms, we skip the frame and move to the next frame. If the delay between the highest incoming depth and color frame number and the currently rendered frame number becomes more than 200 ms, we skip forward and continue with the highest frame number received to avoid latency from individual missing or out of order frames. The transformation information and the US feed are received at 15 FPS and 30 FPS, respectively.

Our distributed 3D scene reconstruction algorithm (4.2) allows the mesh to be constructed from the incoming depth and color frames in real-time on the Hololens. Alternatively, the Hololens view can be rendered on the remote computer and sent with the Holographic Remoting Player [21]. For interaction, the remote Hololens sends the transformation of hands and objects of the remote operator to the local Hololens.

Local HMD Initially, the local Hololens receives the object transformation information from the remote Hololens. Subsequently, it receives a video feed of the remote scene at 30 FPS over a peer-to-peer WebRTC connection. Finally, the US feed and the alignment information from the cameras are received from the local server. The incoming data is used to align and augment the remote information and the local US feed onto the local operator's view.

Remote Computer The remote computer captures the remote expert using a built-in webcam. The video feed is sent at 30 FPS via a WebRTC peer-to-peer connection to the local Hololens. If needed, the remote computer renders the view for the remote Hololens to decrease the load on the HMD. In case the remote video feed is not needed, the remote computer can be removed from the system.

4.2 Distributed 3D Scene Reconstruction

We implemented a distributed 3D dynamic scene reconstruction algorithm. Our lightweight algorithm runs on mobile GPUs and HMDs. The computation is distributed between the GPUs of the local server and the remote HMD. First, the temporal consistency is enforced by the local server. Second, the mesh is spatially smoothed on the remote HMD.

We explain the mesh generation process and our mesh optimization computations for each camera below. To merge the meshes generated by different cameras and to display the views on different physical locations, we apply the alignment computations presented in Section 4.3.

3D Scene Reconstruction per Camera We compute the volumetric representation for each RGBD camera individually. We display the volumetric representation of the local site on a grid of vertices $\boldsymbol{V} \in \mathbb{R}^{288\times320\times3}$. Each vertex on the grid can also be interpreted as a real-world point $P \in \mathbb{R}^3$. For each point $P \in \boldsymbol{V}$ we read the corresponding depth and color value. We use multiple-view geometry [12] to read the correct depth and color values for each grid vertex. Intrinsic and extrinsic camera parameters are provided by the camera software. We provide a detailed description using a similar notation as in the OpenCV documentation [22] in the appendix.

As a result of the color and depth image pixel lookup, we get the 3D position of each vertex on our grid in meters relative to the depth sensor and its color value. We connect neighboring vertices on our grid to create a mesh. However, we only connect vertices less than 0.1 m away from each other. We linearly interpolate the color information from the image onto the mesh.

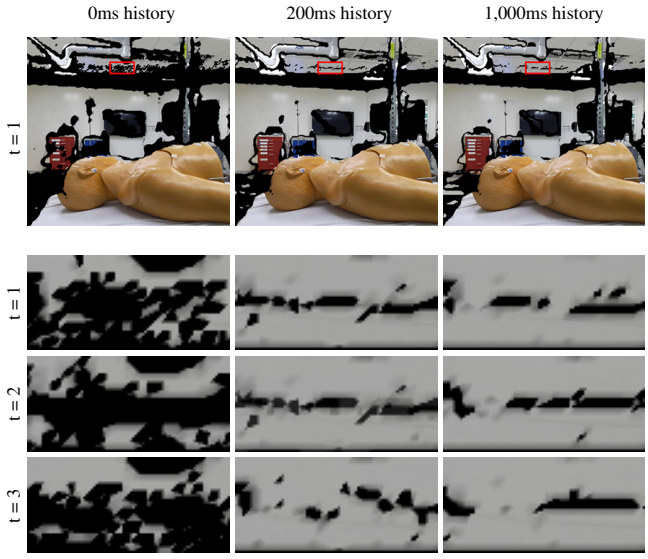


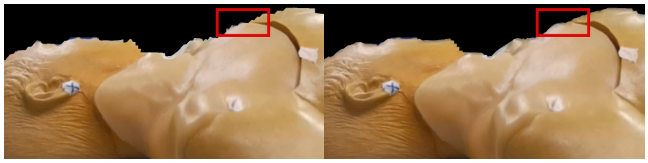
Figure 6: Temporal consistency. We compare three different historic depth sensor reading settings (columns). The top row compares the temporal consistency on a full view. The bottom three rows compare a zoomed in view for three consecutive time steps. Note that increasing the historic reading time windows results in more scene information (top) and less noise (bottom).

Grid Mesh Enhancements The 3D grid mesh we created as explained above has a few visual deficiencies. They are due to the noise of the depth sensor and the nature of the grid topology. Because of the noisy depth data, the created grid mesh is unstable. Vertices move and some of them appear and disappear. Due to the nature of the grid topology, the edges are staircase-shaped. To improve the visual appearance of the 3D mesh, we used a two step process to enhance the mesh. First, we improve the stability of the mesh by temporal smoothing. This stability enhancement is computed on the local server, while maintaining the depth map structure before sending it to the remote Hololens for visualization. Second, we slightly correct the position of edge vertices on the mesh to create smoother object edges. This enhancement is computed on the remote computer that renders the mesh for the remote Hololens.

The following Azure Kinect camera settings provide the best capture quality while keeping bandwidth requirements low for our system. We set the frame rate to 30 FPS and only keep synchronized color-depth image pairs. We set the color resolution to 1920×1080 and the depth resolution to 320×288 using the near field of view (NFOV) 2x2 binned (SW) depth model [20].

We enforce temporal consistency on the local server. We replace invalid sensor readings by the latest historic reading of the last 200ms for each depth pixel at position (i, j). We show an example of how different historic reading times lead to a more stable mesh in Figure 6. We measure the effect quantitatively by capturing one second. 200ms of historic reading recovers 4% of the lost vertices and reduces the on/off vertex flickering by 45%. Increasing the time to 1000ms recovers 7% of the lost vertices and reduces the on/off vertex flickering by 67%. We found that 200ms works best for both static and dynamic scenes.

We tackle small jitter by computing the moving average $\bar{d}$ of the



(a) Grid Mesh (b) Moved + Feathered Edge Mesh

(c) Grid (d) Feather (e) Move (f) Feather + Move

Figure 7: Edge refinement. An example of edge refinement on a mesh side-view of the CVC mannequin. We compare the default grid mesh after combining color and depth information (a) with the refined mesh (b). The individual refining steps are illustrated in detail on a zoomed-in view: grid mesh (c), feathered edges (d), moved vertices (e), and feathered edges combined with moved vertices (f).

valid depth sensor readings $\mathcal{D}(\cdot) > 0$ over the last $n = 10$ frames

$$\bar{d}(i,j) = \Big(\sum_{t=-n}^{-1} \mathcal{D}(i,j,t)\Big) \Big/ \Big(\sum_{t=-n}^{-1} \delta(\mathcal{D}(i,j,t) > 0)\Big), \tag{1}$$

where $\delta(\cdot)$ refers to the indicator function:

$$\delta(\phi(\cdot)) = \begin{cases} 1 & \text{if } \phi(\cdot) \text{ is true} \\ 0 & \text{else.} \end{cases} \tag{2}$$

In case the current depth value $\mathcal{D}(i,j,0)$ is within 3mm of the moving average $\bar{d}$, we assign the previous depth value $\mathcal{D}(i,j,-1)$ to stabilize the vertex. We found $n = 10$ and a 3mm moving average work best for the CVC procedure setup. The small jitter stabilization reduced the mean per frame jitter from 128m to 67m (-48%).

We tackle large jitter by counting the number of changes greater than $\lambda = 3$mm within the last 60 frames $n_2 = 60$,

$$\xi(i,j) = \sum_{t=-n_2}^{-2} \delta(|\mathcal{D}(i,j,t) - \mathcal{D}(i,j,t+1)| > \lambda) \cdot \delta(\mathcal{D}(i,j,t) > 0) \cdot \delta(\mathcal{D}(i,j,t+1) > 0). \tag{3}$$

Similar to small jitter, we assign the previous depth value $\mathcal{D}(i,j,-1)$ if $\xi(i,j)/n_2 > 0.6$. The 0.6 threshold in combination $\lambda = 3$mm and $n_2 = 60$ produced the best results empirically.

The three mesh enhancements explained above (historic reading, small jitter, and large jitter) result in more temporally stable vertices and lower bandwidth requirements. The enhancements combined reduced the mean per frame jitter from 128m to 40m (-68%) in our CVC procedure setup.

After the depth data used to construct the mesh is received by the remote computer, we apply additional mesh enhancements on the vertex level to improve the edge appearance. We show an example of edge refinement in Figure 7. We move the vertices to on-edge positions to remove unnatural edges created by the grid-aligned mesh. Therefore, we consider the 8-neighborhood of adjacent vertices for each vertex. Depending on the number of neighbor vertices within a 10 cm distance, the central vertex is moved in a direction that produces a natural edge. The movement takes into account the grid topology of the mesh. We show our grid topology and examples of how we modify the vertices depending on the number of neighbors in the appendix.

In addition, we set the alpha value of edge vertices depending on their number of vertices. We set the alpha value of every vertex by dividing the number of neighbor vertices within the 10cm distance by 8. This feathering of object edges allows the edges to appear more natural on the remote Hololens.

We only show triangles in the constructed mesh with an edge length smaller than 10cm. This allows us to remove inaccurate information about the scene from a view angle of the scene not captured by the RGBD cameras.

4.3 Positioning and Alignment Between Actors

We align the physical and virtual workspace between the local instructor and the remote expert to enable volumetric communication. First, we focus on the alignment between the two 3D meshes created by the Azure Kinect cameras on the local site. Aligning them allows us to create a volumetric view. Second, we align the volumetric view with the physical world on the local site using point correspondences. The camera and the physical alignment together with the built-in head tracking of the Hololens 2 allows for volumetric communication using pointing, gestures, and virtual tools.

3D View Alignment For finding the rigid transformation $\boldsymbol{R}$ and t between the 3D meshes, we apply least-squares fitting [1] using $N = 4$ point correspondences. Four correspondences result in high alignment accuracy (see Section 5.1) while keeping the setup time low. Least-squares fitting minimizes the error

$$\varepsilon = \sum_{n=1}^{N} ||\boldsymbol{R}\boldsymbol{A}^n + t - \boldsymbol{B}^n||^2, \tag{4}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times 3}$ and $\boldsymbol{B} \in \mathbb{R}^{N\times 3}$ are sets of 3D point correspondences from the 3D meshes. First, we compute the centroids of each point set using

$$\psi(\boldsymbol{S}) = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{S}^n. \tag{5}$$

We subtract the centroid of each point set to center the points around the origin. Then, multiply the two centered point clouds and apply singular value decomposition $SVD(\cdot)$ to find the rotation matrix $\boldsymbol{R}$:

$$\begin{aligned} (\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}) &= SVD((\boldsymbol{A} - \psi(\boldsymbol{A}))(\boldsymbol{B} - \psi(\boldsymbol{B}))^T) \\ \boldsymbol{R} &= \boldsymbol{V}\boldsymbol{U}^T. \end{aligned} \tag{6}$$

Once we checked for reflection $|\boldsymbol{R}| < 0$, we compute the translation

$$t = \psi(\boldsymbol{B}) - \boldsymbol{R}\psi(\boldsymbol{A}). \tag{7}$$

The resulting rotation $\boldsymbol{R}$ and translation t describe the rigid transformation between different camera or world views given 3D point correspondences $\boldsymbol{A}$ and $\boldsymbol{B}$.

We use the transformation computation presented above for aligning the two camera views on the local site to create a volumetric representation at the remote site. Moreover, we compute the rigid transformation between the physical local Hololens device and camera 1 to align remote gestures and objects to the physical world of the local operator.

Gestural Communication We built our AR application for the Hololens 2 HMDs using the Mixed Reality Toolkit. The toolkit allows us to support standardized AR user interaction. On the remote instructor's Hololens, we used Mixed Reality Toolkit's pose detection to predict the hand position. The head-tracking is also used for alignment between the nodes.

For gestural communication, we send a hand model representing the remote expert's hands to the local operator. We detect the hand gestures using the camera system of the Hololens 2. The 3D position

and rotation of 26 hand joints are sent to the local operator. We send the hand joint data $\{R,t\}_1$ in camera 1 coordinates:

$$\begin{aligned} \boldsymbol{R}_1 &= \boldsymbol{R}_2^{-1}\boldsymbol{R}_3, \\ t_1 &= \boldsymbol{R}_2^{-1}(t_3 - t_2), \end{aligned} \tag{8}$$

where $\{R,t\}_2$ refers to the remote volumetric view transform, and $\{R,t\}_3$ refers to the hand joints in remote Hololens coordinates. At the local site a hand model is animated to represent the remote instructor's gestures using Equation (8) where $\{R,t\}_1$ represents the hand joints in local Hololens coordinates, $\{R,t\}_2$ represents the local to camera 1 transformation, and $\{R,t\}_3$ represents the hand joints in camera 1 coordinates. We estimate $\{R,t\}_2$ using the point correspondence optimization illustrated in Equation (4). We apply the same transformation for virtual objects sent from remote to local.

Initial Setup Procedure We propose an initial calibration phase during system setup to align the RGBD camera coordinate systems and the Hololens coordinate systems. For convenience, 4 markers are stuck on a static object in the local scene, close to the area of interest. Note that any landmark can be used instead of markers for this step. Both RGBD cameras see the markers. The local operator then places virtual points on top of the markers for each camera on the local server AR application. The local Hololens camera system is also calibrated using the markers. The references between physical marker, Hololens, and Azure Kinect camera 1 enable volumetric collaboration.

The virtual workspace setup on the local Hololens is relative to the physical markers as shown in Figure 5. The remote expert initiates the workspace standing behind the desired workspace location. The virtual workspace appears in front of them similar to a monitor setup as shown in Figure 4. The remote expert then fine-tunes the position and rotation of the volumetric view using hand gestures.

5 Evaluation

We evaluated the proposed mixed reality system in a study in which medical experts taught the ultrasound-guided placement of a central venous catheter to learners. We conducted the study in a simulation center using CAE Blue Phantom ultrasound central venous access training mannequins [30]. We compared training with the proposed MR system against training with video communication software. For video communication, we capture the same views as the MR system, a side, and an over-the-shoulder view. The results were analyzed using surveys and video recordings.

5.1 System Setup Analysis

We constructed a modular camera mount to allow for flexible camera positioning. Yet, the camera mount is rigid and provides stable camera views throughout the procedure. Our camera mount is attached to the stretcher using existing bolts. This allows for fast setup and consistent camera positions relative to the mannequin between sessions. Moreover, we added a pin mount to the base of the camera mount to secure the mannequin position on the stretcher. After we evaluated different views with domain experts, we found that a side combined with a top camera position provides the best view for the procedure. We mounted the side camera and the top camera at distances of 86cm and 103cm, respectively, relative to the laryngeal prominence of the mannequin. Both cameras were rotated such that they point to the area relevant for the US-CVC procedure.

We measured the alignment accuracy between the local and remote after the system setup. Each time, we took four measurements evenly distributed around the borders of the tissue insert from the CVC mannequin, at the critical area of the procedure. We found the mean error between the remote operator's volumetric view and the local physical to be 1.36cm, $\sigma = 0.18$cm, for $n = 10$ setups. The participants of our study reported that the accuracy is sufficient for giving pointing instructions and visual object guiding. The alignment error after the initial setup between the two camera views on the remote side is 2.54cm, $\sigma = 0.34$cm, for $n = 5$ setups. We lower the initial in-between camera error manually. This is possible because of our static camera setting on the local site.

Age	26.9y
Male/Female	6/14
Clinical training and/or practice experience	2y
Prior AR/MR/VR experience	0

Table 1: Learner demographics and prior experience.

5.2 Study Participants

Our study participants consisted of a group of 5 instructors and 20 learners, all lived and trained in the USA. We randomly assigned the instructors and the learners to 10 mixed reality and 10 video training sessions. Each learner completed exactly one training session whereas instructors taught multiple sessions. Each instructor taught at least one video and one mixed reality session. We illustrate the learner demographics and prior experience in Table 1. The instructors were on average 43 years old and consisted of four males and one female. All of them were performing US-CVC for more than 3 years and teaching the procedure for more than 1 year.

5.3 Study Setup

We prepared a Blue Phantom ultrasound central venous access training mannequin [30], a CVC kit, and a Sonosite M-Turbo Ultrasound system [34]. The following parts of the CVC procedure were taught:

1. A talk through the procedural steps, the preparation of the CVC kit, and the use of the ultrasound.

2. Catheter placement over a wire using the Seldinger technique facilitated through the catheter over the needle approach. Confirmation of wire placement is achieved with ultrasound.

3. The flushing and drawing of the three catheter ports after insertion.

We compared training with the proposed MR system against training with video communication software. For video communication, we capture the same views as the MR system, a side, and an over-the-shoulder view.

5.4 Study Process

Prior to the study, learners and instructors provided informed consent. The learners completed US-CVC pre-training to familiarize themselves with the steps of the procedure. Instructors and learners had not received MR training prior to the study. The learners completed a pre-training survey which included demographic and prior experience questions. At the beginning of each training session, instructors talked about background information on US-CVC with the learner. Then, they prepared the learner's workspace for the procedure and talked through the medical equipment necessary for the procedure. After the initial preparation, instructors moved to a separate room to start with the video or mixed reality training.

In the case of MR sessions, both instructors and learners completed a 5-minute MR briefing. Apart from the briefing, the participants did not receive any training on the technology. Then, the actual training session, which took about one hour, started. After the US-CVC session, both learners and instructors completed surveys and interviews.

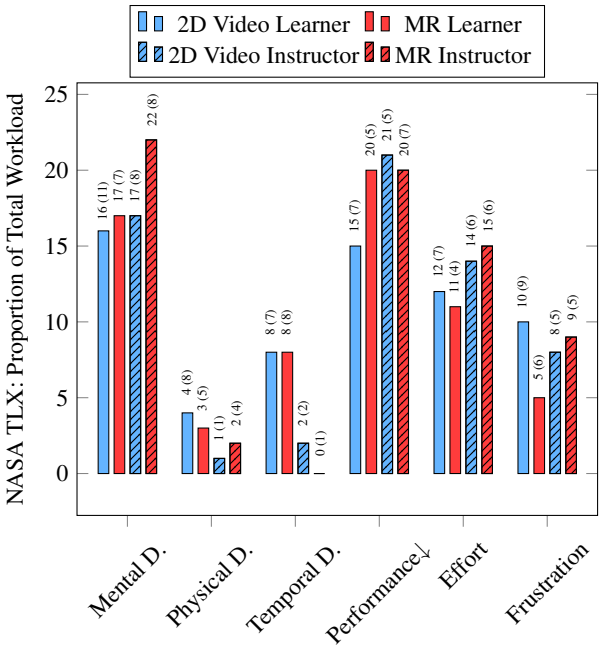


Figure 8: NASA TLX workload. We compare the weighted NASA TLX score (y-axis) per category (x-axis) for learner and instructor using video and MR teaching. We represent the video and MR results using blue and red bars, respectively. The instructor bars are filled with a line pattern. The standard deviation is shown in parenthesis.

5.5 Study Results

We analyzed post-session interviews, recorded video data, and NASA-TLX survey responses to evaluate the mixed reality system. Overall, the feedback from both instructors and learners was very positive on using the mixed reality communication system for US CVC training.

Workload Analysis The NASA-TLX gave us a quantitative subjective measure of the workload for instructors and learners. The validated instrument allowed us to compare video and MR sessions, see Figure 8. Our null hypothesis was that video and MR training results in equal workload. We could not reject our null hypothesis in total as well as in per category workload between MR and video by performing a two-sample two-tailed t-test using a significance level $\alpha = 0.05$. This, in HCI commonly chosen significance level, is important to minimize the uncertainty.

Instructor Feedback From interviews and observation of the instructor, we learned that each volumetric view, US view, and 2D view are essential during different parts of the procedure and for different purposes. The instructors liked the volumetric view because it gave them a spatial understanding of the scene and it allowed for visual communication using gestures and objects. However, smaller and translucent surfaces were sometimes not captured correctly in the volumetric view making it difficult for instructors to identify them. To overcome this issue, the instructors used the video feed as a backup. We also observed that virtual objects in combination with gestural communication and the volumetric view can be used to effectively teach the correct usage of medical equipment. This turned out to be especially useful for teaching needle-probe coordination. The US feed together with the volumetric view gave the remote instructor a good spatial understanding of the needle-probe guidance of the local learner during an essential part of the procedure.

Learner Feedback The learners reported that the augmented instructor's hands and objects helped them learn how to use the medical equipment much faster than using verbal instruction only. However, the learners also reported that the teachers augmented hands sometimes were distracting because they were visible throughout the procedure and it was not clear if they are actively used by the instructor for communication. This suggests minor modifications in the software. Moreover, the learners highlighted the importance of the instructor's webcam view to see who they are communicating with. The opinions on the augmented US feed were mixed. Some learners liked it because it was positioned very close to the procedure area. Others preferred to use the physical US screen.

5.6 Discussion

The fact that the subjective workload was not significantly higher is considered a positive result for the design of the system given that it was the first time that the learners used XR (AR, MR, VR) and the Hololens 2. Because video conferencing technology is used in everyday life it is expected to produce a lower extraneous workload. Although the instructors were familiar with the basic functionality of the system before they first used it, they did not have full training with the system before their first study session. Thus, we argue that our system is very intuitive to use and it only requires a five-minute hardware and user interface debrief before using it.

We observed a learning effect of the instructor teaching through MR. The more often they taught, the more they utilized virtual objects and hand gestures. Hence, MR with experienced users might result in a lower workload and a better experience.

Analyzing the individual workload categories, we argue that the trend toward higher frustration for the learner using video came from the fact that complex parts of the procedure were harder to understand using 2D visuals and needed multiple iterations of explaining. A reason for the high mental demand for the instructor can be that mixed reality has many options to teach and observe.

6 Conclusion

We presented a mixed reality (MR) real-time communication system for assistance during ultrasound-guided central line placement. The system allows a remote expert to connect to a local operator to guide them in MR. The MR interaction allows for vocal and gestural communication. The communication is based on volumetric capture through RGBD cameras that allow for intuitive visual guidance. We proposed an algorithm that focuses on lightweight, real-time communication and rendering of volumetric capture.

We evaluated the proposed system in a user study in which we compare MR against video communication for ultrasound-guided CVC placement training. We found that MR provides a viable alternative compared to video during the CVC procedure training. We showed how the different elements of the system can be used effectively during procedural training.

While we focus on a single medical procedure, the issue of training and assistance by an expert who is not on-site is present across multiple disciplines and many domains which depend on operator cognitive and manual procedural skills [9, 28]. When mastery-level skill must be brought to a remote location during natural disasters, epidemics, equipment breakdowns, etc., our system could be sent to the remote location or travel with providers who may be in need of support.

Acknowledgments

The work is supported by National Science Foundation grant no. 2026505 and 2026568. The authors wish to thank Erin Horan, Safinaz Alshiakh, Yasser Ajabnoor, Ahmed Allabban, Becky Lake, Scott Schechtman, Rahil Ashraf, and Carine Cristina Goncalves Galvao for their help. Moreover, the authors would like to thank medical students and residents for participating in the experiment.

REFERENCES

[1] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9:698–700, 1987.

[2] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies. Neural rgb-d surface reconstruction. *arXiv preprint arXiv:2104.04532*, 2021.

[3] A. Bozic, M. Zollhofer, C. Theobalt, and M. Niessner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] A. X. Chang, A. Dai, T. A. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from RGB-D data in indoor environments. *CoRR*, abs/1709.06158, 2017.

[5] J. Chen, D. Bautembach, and S. Izadi. Scalable real-time volumetric surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(4):1–16, 2013.

[6] M. C. Davis, D. D. Can, J. Pindrik, B. G. Rocque, and J. M. Johnston. Virtual interactive presence in global surgical education: international collaboration through augmented reality. *World Neurosurgery*, pp. 103–111, 2016.

[7] D. Gasques, J. G. Johnson, T. Sharkey, Y. Feng, R. Wang, Z. R. Xu, E. Zavala, Y. Zhang, W. Xie, X. Zhang, K. Davis, M. Yip, and N. Weibel. Artemis: A collaborative mixed-reality system for immersive surgical telementoring. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[8] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Trans. Graph.*, 36(4), jun 2017. doi: 10.1145/3072959.3083722

[9] E. Hadar, J. Shtok, B. Cohen, Y. Tzur, and L. Karlinsky. Hybrid remote expert - an emerging pattern of industrial remote support. In *CAiSE-Forum-DC*, 2017.

[10] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1524–1531, 2014. doi: 10.1109/ICRA.2014.6907054

[11] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, pp. 904–908, 2006.

[12] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[13] E. S. Holmboe, L. Edgar, and S. Hamstra. The milestones guidebook. *Chicago, IL: Accreditation Council for Graduate Medical Education*, 2016.

[14] J. Kessler, J. T. Wegener, M. W. Hollmann, and M. F. Stevens. Teaching concepts in ultrasound-guided regional anesthesia. *Current Opinion in Anaesthesiology*, 29:608–613, 2016. doi: 10.1097/ACO.0000000000000381

[15] J. Kim, H. Kim, H. Nam, J. Park, and S. Lee. Textureme: High-quality textured scene reconstruction in real time. *ACM Trans. Graph.*, 41(3):24:1–24:18, 2022. doi: 10.1145/3503926

[16] S. Kim and J. Kim. Occupancy mapping and surface reconstruction using local gaussian processes with kinect sensors. *IEEE Transactions on Cybernetics*, 43(5):1335–1346, 2013. doi: 10.1109/TCYB.2013.2272592

[17] C. Legoux, R. Gerein, K. Boutis, N. Barrowman, and A. Plint. Retention of critical procedural skills after simulation training: A systematic review. *AEM Education and Training*, 5(3):e10536, 2021. doi: 10.1002/aet2.10536

[18] F. Mahmood, E. Mahmood, R. G. Dorfman, J. Mitchell, F. U. Mahmood, S. B. Jones, and R. Matyal. Augmented reality and ultrasound education: Initial experience. *Journal of Cardiothoracic and Vascular Anesthesia*, 32:1363–1367, 6 2018. doi: 10.1053/j.jvca.2017.12.006

[19] S. Meerits, V. Nozick, and H. Saito. Real-time scene reconstruction and triangle mesh generation using multiple rgb-d cameras. *Journal of Real-Time Image Processing*, 16(6):2247–2259, 2019.

[20] Microsoft. Azure kinect dk hardware specifications, May 2022. Available at `https://docs.microsoft.com/en-us/azure/kinect-dk/hardware-specification`.

[21] Microsoft. Holographic remoting player overview, May 2022. Available at `https://docs.microsoft.com/en-us/windows/mixed-reality/develop/native/holographic-remoting-player`.

[22] OpenCV. Camera calibration and 3d reconstruction, May 2022. Available at `https://docs.opencv.org/3.4.1/d9/d0c/group__calib3d.html#ga13f7e34de8fa516a686a56af1196247f`.

[23] K. Pietroszek. Virtual hand metaphor in virtual reality. *Encyclopedia of Computer Graphics and Games*, pp. 1–3, 2018.

[24] K. Pietroszek and C. C. Lin. Univresity: Face-to-face class participation for remote students using virtual reality. In *25th ACM symposium on virtual reality software and technology*, pp. 1–2, 2019.

[25] H. Prescher, E. Grover, J. Mosier, U. Stolz, D. Biffar, A. Hamilton, and J. Sakles. Telepresent intubation supervision is as effective as in-person supervision of procedurally naive operators. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association*, 21, 12 2014. doi: 10.1089/tmj.2014.0090

[26] M. Rebol, C. Gütl, and K. Pietroszek. Passing a non-verbal turing test: Evaluating gesture animations generated from speech. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 573–581. IEEE, 2021.

[27] M. Rebol, C. Gütl, and K. Pietroszek. Real-time gesture animation generation from speech for virtual human interaction. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411763.3451554

[28] M. Rebol, C. Hood, C. Ranniger, A. Rutenberg, N. Sikka, E. M. Horan, C. Gütl, and K. Pietroszek. Remote assistance with mixed reality for procedural tasks. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 653–654, 2021.

[29] E. Rojas-Muñoz, M. E. Cabrera, D. Andersen, V. Popescu, S. Marley, B. Mullis, B. Zarzaur, and J. Wachs. Surgical telementoring without encumbrance: a comparative study of see through augmented reality based approaches. *Ann Surg*, pp. 384–389, 2019.

[30] W. science. Blue phantom® central line ultrasound trainers, May 2022.

[31] M. B. Shenai, R. S. Tubbs, B. L. Guthrie, and A. A. Cohen-Gadol. Virtual interactive presence for real-time, long-distance surgical collaboration during complex mucrosurgical procedures. *J Neurosurg*, pp. 277–284, 2014.

[32] W.-X. Si, X.-Y. Liao, Y.-L. Qian, H.-T. Sun, X.-D. Chen, Q. Wang, and P. A. Heng. Assessing performance of augmented reality-based neurosurgical training. *Visual Computing for Industry, Biomedicine, and Art*, p. 6, 2019.

[33] B. Sites, B. Spence, J. Gallagher, C. Wiley, M. Bertrand, and G. Blike. Sites bd, spence bc, gallagher jd, wiley cw, bertrand ml, blike gt. characterizing novice behavior associated with learning ultrasound-guided peripheral regional anesthesia. *Regional anesthesia and pain medicine*, 32:107–15, 03 2007. doi: 10.1016/j.rapm.2006.11.006

[34] Sonosite. Sonosite m-turbo, May 2022.

[35] S. Wang, Y. Kwon, Y. Shen, Q. Zhang, A. State, J.-B. Huang, and H. Fuchs. Learning dynamic view synthesis with few rgbd cameras. *arXiv preprint arXiv:2204.10477*, 2022.

[36] A. Wickey and L. Alem. Analysis of hand gestures in remote collaboration: Some design recommendations. In *Proceedings of the 19th Australasian Conference on Computer-Human Interaction: Entertaining User Interfaces*, p. 87–93, 2007.

[37] Y.-S. Wong, C. Li, M. Nießner, and N. J. Mitra. Rigidfusion: Rgb-d scene reconstruction with rigidly-moving objects. *Computer Graphics Forum*, 40(2):511–522, 2021. doi: 10.1111/cgf.142651

[38] H. Xu, X. Wang, and L. Shi. Fast 3d-object modeling with kinect and rotation platform. In *2015 Third International Conference on Robot, Vision and Signal Processing (RVSP)*, pp. 43–46, 2015. doi: 10.1109/RVSP.2015.19

[39] T. Yu, Z. Zheng, K. Guo, P. Liu, Q. Dai, and Y. Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5746–5756, June 2021.