EditAR: A Digital Twin Authoring Environment for Creation of AR/VR and Video Instructions from a Single Demonstration

Subramanian Chidambaram* Purdue University
Sai Swarup Reddy† Purdue University
Matthew Rumple‡ Purdue University
Ananya Ipsita§ Purdue University
Ana Villanueva¶ Purdue University
Thomas Redick‖ Purdue University
Wolfgang Stuerzlinger** Simon Fraser University
Karthik Ramani†† Purdue University

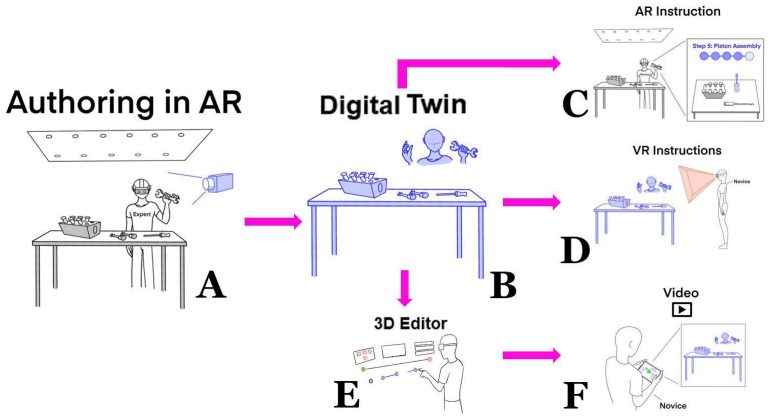

Figure 1: Overview of the EditAR authoring. (A) Expert uses physical tools to create instructions. (B) A digital twin of the user's action is captured. (C) Export of AR instructions for use in a similar environment. (D) Export of VR instructions for viewing in an immersive VR environment. (E) A 3D Editor enabling Subject Experts to create 2D content from the digital twin recording. (F) Export of video instructions for viewing on a traditional display.

Abstract

Augmented/Virtual reality and video-based media play a vital role in the digital learning revolution to train novices in spatial tasks. However, creating content for these different media requires expertise in several fields. We present EditAR, a unified authoring, and editing environment to create content for AR, VR, and video based on a single demonstration. EditAR captures the user's interaction within an environment and creates a digital twin, enabling users without programming backgrounds to develop content. We conducted formative interviews with both subject and media experts to design the system. The prototype was developed and reviewed by experts. We also performed a user study comparing traditional video creation with 2D video creation from 3D recordings, via a 3D editor, which uses freehand interaction for in-headset editing. Users took 5 times less time to record instructions and preferred EditAR, along with giving significantly higher usability scores.

*e-mail: schidamb@purdue.edu
†e-mail: reddy37@purdue.edu
‡e-mail: rumple0@purdue.edu
§e-mail: aipsita@purdue.edu
¶e-mail: villana@purdue.edu
‖e-mail: tredick@purdue.edu
**e-mail: w.s@sfu.ca
††e-mail: ramani@purdue.edu

1 Introduction

Kinesthetic or hands-on learning is defined as "the ability to process information through touch and movement," [51]. Kinesthetic learning is integral to skill acquisition for spatial tasks, including welding [21, 30], machining, assembly, home repair, and surgery [27]. Currently, one-on-one instruction, video-based, and paper-based methods are widely used instructional transfer modes for kinesthetic skill acquisition. However, there is growing evidence that Augmented and Virtual Reality (AR/VR) instructions improve task completion rate, reduce error rates, and provide better usability, as demonstrated by Ipsita et al. [21], Henderson et al. [16], Marner et al. [29], Mohr et al. [33], and YouMove [1].

In the past, video-based platforms, such as SkillShare [52], Instructables [20], and YouTube [69] have all seen exponential growth in instructional content. All these platforms are video-based, and creators are quite prolific with this "default" authoring medium. However, recent years have produced substantial advances in low-cost, untethered headsets, such as the Oculus Quest 2 [40], and commercial interest in immersive media is on the rise [24]. Thus, we believe there is a need for developing and studying immersive media authoring environments, in AR or VR, while (a) still supporting widely-used "default" authoring multimedia such as video, and (b) enabling subject matter experts without multimedia/programming skills to participate in this creation process. Such tools will enable the expansion and wide adoption of these new instructional technologies.

While having multiple options for multimedia consumption allows people with different hardware platforms and resources to participate in learning activities, it also requires instructors to possess multiple skill-sets to author content for multiple media. That

978-1-6654-5325-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ISMAR55827.2022.00048

is, to develop content for all authoring these media and alongside their subject expertise, Subject Matter Experts (SMEs) are expected to know 3D modeling, video editing, AR/VR programming, and instructional design. The alternative is for SMEs to work with different professionals to create content, leading to increased resource requirements in terms of cost and effort. Additionally, the need for multiple demonstrations of the same task, where the SME demonstrates the same task multiple times to support content creation for different media, increases time requirements substantially.

Recent work in immersive technology has demonstrated successful capture and replay of reality, effectively making reality "asynchronous" [9]. We leverage this approach to capture reality and to support instructional multimedia creation for AR, VR, and video-based media, while only requiring a single task demonstration without relying on multimedia expertise. This work presents EditAR, a digital twin-based authoring and editing environment. We define a digital twin as an executable virtual model of a physical thing or a system [18, 66]. Through external 6DoF tracking sensors, EditAR captures and creates a digital twin of user interaction within the environment, without relying on hand-held controllers.

EditAR represents all objects/tools within the environment of interest in a virtual model and the user is represented as a virtual avatar. A digital twin recording (also known as a 3D recording) is a virtual volumetric video consisting of a 3D scene with a temporal component. EditAR uses this recording to later enable the creation of AR and VR content. The immersive nature of both AR and VR allows the content to be simply replayed from the author's demonstration over the physical or virtual worlds. In contrast, creating 2D videos from a virtual 3D recording requires a unique editing environment. To enable users to capture 2D video from the digital twin recordings with the aid of virtual cameras, EditAR offers a 3D video editing interface. Based on formative interviews with media creation experts, we evaluated several variations in a preliminary study and selected three such virtual cameras for inclusion in EditAR. Finally, we evaluated EditAR as a whole with AR, VR, and video experts to ascertain the system's effectiveness. Our main contributions are:

- A unified and efficient system that empowers SMEs to author and edit content for AR, VR, and video-based media. The authoring requires only a single demonstration and no multimedia expertise.
- A 3D editing interface for free-hand, in-headset interaction to create 2D videos from 3D recordings by exploring novel virtual camera interactions, which were created based on media expert input, followed by a preliminary study to verify interface design decisions.
- An expert review of the EditAR and its interface to evaluate the system's effectiveness.
- A study with novice users comparing traditional 2D video creation and our 3D editing interface that offers free-hand, in-headset interaction to create 2D videos from 3D recordings.

2 Related Work

2.1 AR authoring environments

With affordable Head-Mounted Displays (HMDs), tablets, and smartphones, access to AR and VR is becoming more commonplace. Thus, recent work to enable non-programming experts to create immersive content, such as AuthAR [65] and ProcessAR [7], proposed codeless AR authoring techniques. Current commercial AR authoring tools such as Vuforia Chalk [46], PTC's Vuforia Capture [45], and Teamviewer pilot [56] either render 2D annotations over the physical space or capture 2D video to be embedded later over the real world. Yet, they provide no support for interactive authoring with immersive 3D content. While Microsoft 365 Dynamics [31] supports rendering 2D and 3D content, this requires post-processing via a desktop keyboard and mouse interface, similar to Unity 3D. While such work makes authoring content somewhat more accessible, it typically targets only a single kind of multimedia. Also, such approaches often require creators to edit and re-create content with other systems (e.g., video editing for a VR recording), requiring additional skills to create multimedia. As the diversity in multimedia types and demand to support multiple platforms increases, current workflows fall short, which is addressed by EditAR.

Past work such as RealitySketch [55] and Pronto [25] allowed users to sketch over physically captured video. While RealitySketch [55] enabled users to visualize real-world phenomena through responsive graph plots and interactive visualizations, Pronto [25] allowed users to prototype dynamic AR designs quickly. While these works are excellent examples for AR applications and provide AR authoring platforms, they do not address the growing need to support several kinds of multimedia such as AR, VR, and video.

2.2 AR through Motion Capture

Ong et al. [42], Radkowski [48], and Funk et al. [10] presented workflows that allow the user to use their bare hands to interact with virtual objects. While these approaches preserve the natural and intuitive interactions that we desire to record, these authoring environments are limited due to the lack of haptic feedback while authoring high-fidelity instructions. In contrast, ARtalet [14] emphasized the need for such haptic feedback while authoring AR instructions, but required specialized manipulation props that the instructors had to build first. EditAR provides the necessary passive haptic feedback simply through real tool interaction, tracked via 6Dof sensors and without requiring specialized props.

To capture user motions, Loki [57] and Oda et al. [41] used the Azure Kinect and Optitrack [43] technology, respectively. Zillner et al. [70] provided dense scene reconstruction with 3D meshes to facilitate remote instructional delivery but explicitly targeted synchronous media. Thus, a constant presence of an expert instructor is required for active learning. While this approach has benefits, the design principles behind this approach are not directly applicable for *asynchronous* authoring systems such as EditAR.

2.3 Record and Replay in 3D/VR Environment

Recording and replaying within a 3D environment have been explored in games such as Starcraft II and Toribash [58]. Galvane et al. [11] and XR studio [36] provided pre-visualization tools and virtual production techniques to support film-making and live immersive streaming instructions for VR and MR, respectively. Vreal [62] is a plugin for VR games that allowed users to record and store their VR content. Earlier work such as "Just follow me" [68] and recent work such as "Again, together" [63] have explored recording and replaying of 3D virtual avatars. The Virtual Mail system [19] supported recording an avatar's gestures and audio together with the surrounding environment. vAcademia [34] allowed users to record and share their presentations in a 3D virtual environment. "Who put that there?" [26] allowed for temporal navigation within a spatial recording to track the position and location of various virtual objects.

As most of these works relied only on an immersive environment, there was no interaction with the real world. In contrast, our work focuses on kinesthetic instructions, which involve direct interactions with real world objects. Also, most of the earlier research mentioned above focused on capturing content and re-viewing in an HMD [26, 63], which prevents traditional media users from consuming content. On the other hand, based only on a single demonstration of the recorded 3D scenario, EditAR instead offers novel virtual camera interactions to produce video instructions. Other past 3D record-and-replay systems also did not leverage the strength of asynchronicity to use a single demonstration to produce multiple types of content.

2.4 In-Headset Editing

Vremiere [37] presented a VR headset-based interface for editing monocular panoramic videos with a keyboard and mouse. Recent work such as 6Dive [13] provided a similar editing interface with 6DoF controllers. While both works explored the concept of immersive video editing, they cannot be directly applied to suit the needs of EditAR. As EditAR's users need to interact with physical tools and objects, they are not able to remain seated in front of a desk, nor can they interact with controllers. A hands-free, gesture-based editing interface is required, which EditAR explores.

Finally, commercially available tools such as nvrmind [38], Tvori [60], and quill.art [47] support in-headset editing and creation of 3D animation via interactive keyframes and the capture and editing of the virtual video within an HMD through the controller. But like past research, these commercial tools fall short, as there is a stark disconnect between the physical and digital world within the headset. That is, they do not enable the capture of physical work interactions within their workflow.

3 Design Space Exploration

After exploring the research literature, we interviewed twenty experts, including SMEs educators and multimedia experts, such as videographers, AR, and VR programmers. We identified real-world constraints and design suggestions for an AR system that can help with the large-scale deployment of instructional authoring.

3.1 Formative Interview

Our twenty interviewees included eight subject experts and three educators in manufacturing, five AR/VR programmers, and four videographers with film and VFX backgrounds. The interview began with the experts sharing their experiences, current constraints, and practices in their industry to train new employees, and what they expect from a product to author instructions. We distilled the most relevant information from these interviews and present it here, as a design rationale for the features in EditAR.

There was consensus among the experts that there are multiple types of instructional media, each with its strengths and weaknesses. The SMEs also realize that newer technologies such as AR and VR are being explored, but they were unsure where they would fit in their training programs and how to create and use such content. The AR and VR media experts, also identified that the content they create has not reached their target audience as much as 'video' has and that they are limited by the hardware available to the end user. Still, they recognized that their reach is increasing as cheaper hardware becomes available. Thus, we identified a primary design objective for EditAR to be the ability to create instructional content for multiple media, in order to accommodate end users with different available hardware.

The SMEs emphasized that spatial tasks such as machine operation or repair require a holistic understanding of all components of the machine. Videographers said they usually capture the action from multiple angles and edit the result to provide a complete overview for the learner. They also stated that this approach requires numerous demonstrations of the same action seen from different points of view, similar to immersivePOV's findings [17]. Hence, we realized a need for "capturing multiple perspectives." Finally, all experts emphasized the need for having the "human-in-the-loop" during the editing process, as they felt that a completely automated system might not be able to anticipate the needs of different learners.

The video experts' interviews provided vital insights for developing the 3D editing interface to capture 2D video content. The video experts described different kinds of cameras and filming techniques that are currently used. Based on their description, we divided and categorized the types of cameras into two broad categories: Dynamic and Static. Dynamic cameras move in space to capture the environment from different perspectives over time. Static cameras remain stationary in space during capture. Depending on the input received to move the dynamic cameras, they can be further classified into automated or manual cameras. The inference from the discussions was that each category of cameras has its purpose, and depending on the situation, different categories are viable.

To avoid overloading the user with choices while still maximizing the utility of our work, we conducted a preliminary study to evaluate a set of 10 different virtual cameras, developed by studying current physical camera equipment and techniques. We used products such as tripods, Camera Slider Rigs [28], Skycam [8], Gimbals [49], and techniques, such as television sports production [44], as design inspirations for creating the virtual camera metaphors. We also incorporated comments from SMEs and education content creators, emphasizing the need to focus on the user's hands or to provide a first and third-person perspective of the work environment to develop other camera metaphors. The details of this study are described below in section 4.2.3.

4 Overview

EditAR actively tracks the position and orientation of objects in the scene and allows the user to author content by simply demonstrating the task. In addition to object information, EditAR also captures the position and orientation of the user's hands and upper body to create an equivalent digital twin that can later be used to create immersive instructions directly for AR and VR. We decided to use a half-body avatar for EditAR, as an exploratory study by Cao et al. [6] identified that most users preferred a half-body avatar representation to other ones, such as a full-body avatar, non-avatar, or direct video tutorials. To support the creation of 2D videos, EditAR also provides an editing interface that uses virtual cameras and the digital twin.

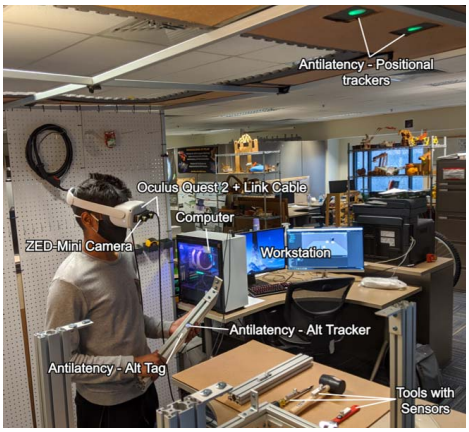

Figure 2: Hardware setup for system implementation. The modified Oculus Quest 2 VR head-mounted display supports video pass-through AR with a ZED mini depth camera.

4.1 Architecture & Hardware

EditAR was developed and deployed on an AR pass-through system by attaching a stereo camera (ZED mini Dual 4MP camera with 2560x720 resolution [53]) to an Oculus Quest 2 [40] VR headset. The Quest 2 was connected to a PC (3.8 GHz AMD Ryzen 7 5800X, 32GB RAM, nvidia GTX 1660) via a 'link' cable. Our tracking system uses a 7x7x7ft (2.1x2.1x2.1m) aluminium structure made of 80/20 Quick Frame. For 6DoF tool/object tracking, we use Antilatency's system [2] providing a 10x10x10ft (3x3x3m) tracking area, with 'Alt Tags' and 'Alt Trackers' tracking modules. The sensor data was wirelessly transmitted to the PC via Antilatency's 'HMD Radio Sockets' [2]. To minimize unwanted occlusions, the tracking system uses 12 active markers on the ceiling to determine the tracker positions (refer Fig 2). EditAR was developed in Unity 3D

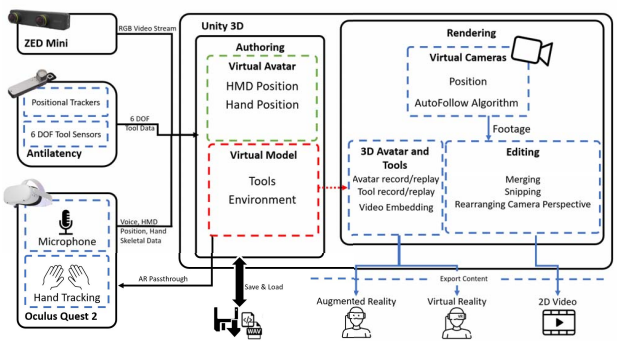

Figure 3: System Architecture: Overview of the data flow from the different hardware used for various system sub-models.

(2019.4.21f). We relied on the Oculus SDK [39] for avatar visualization and recording. We used the standard Oculus SDK hand prefabs and API for hand tracking and visualization.

4.2 Walkthrough: Capturing instructions for assembling an 80/20 structure

We use the example of assembling an 80/20 rig structure to present a sequence of steps users follow to create content using EditAR. We describe the various features of EditAR throughout the walkthrough.

4.2.1 Step 1: Preparing for Authoring

Before authoring, the user first defines custom hand gestures and assigns them to invoke the features of their choice.

Hand Gesture Calibration. Using Oculus Quest 2's hand tracking API, we allow users to create and utilize custom hand gestures within our application. Based on work such as GesturAR [64], which highlighted the importance of hand dexterity and robust hand gesture detection, we decided to give the users the freedom to determine their own gestures for any function. The user can define a gesture by performing the desired gesture with one hand and tapping 'Calibrate Gesture' with their other hand. Our gesture recognition algorithm uses the relative vectors of the nodal hand joints with respect to the wrist position to recognize gestures. For identifying the gesture, the current nodal positions are then compared against the calibrated values with a threshold.

Virtual-Physical Alignment. One of EditAR's strengths is its ability to produce a Digital Twin of real-world physical interactions. For this, the author needs to identify all relevant objects within the workspace and align the physical and virtual models (widely available in the form of CAD models and 3D assets [22, 50, 54, 59]) before authoring. Objects within a workspace can be categorized as either static environmental objects or dynamic objects such as tools and equipment. To prepare the physical objects involved in the process, the user attaches an Antilatency tracking module to all objects or tools of interest. In the case of the 80/20 rig, all the components that would move during the assembly process are tagged. After attaching the tracking modules, the user only needs to align a physical tool with the corresponding virtual model. The virtual model is loaded through EditAR's 3D UI element and aligned in AR to the physical tool for calibration. Later, after successfully capturing action or a whole sequence, the author ends/concludes a recording by invoking another UI element through a user-defined gesture.

4.2.2 Step 2: Authoring

EditAR allows the user to author instructional content by simply performing the task at hand. In the case of the 80/20 rig, the user assembles and interacts with the objects in question as they usually would. EditAR records a corresponding Digital Twin in the background that captures the movement of all the equipment and the user's hands and head/body movements in 3D space.

Record and replay for tools and avatar in the 3D scene. The recording function is automatically triggered once the user interacts with a tool. For each moment, we record a *Tracking Frame*, which contains timestamped information about the current poses of the avatar, hands, finger joints, and all tools and equipment of interest. All such Tracking Frames are stored in a buffer that is serialized using Unity3D's serialization library for easy saving of recordings. The de-serialization protocol from the same library is used to load a recording to be replayed or scrubbed through. During replay, the timestamps stored in the Tracking Frames regulate the playback speed of the recordings, allowing the recording to be played back at the exact speed it was recorded, regardless of the system framerate.

2D Video Embedding. Similar to previous work [7, 31, 45, 65] we prefer to use 2D video to convey instructions for fine-grained interactions, like inserting a screw into the corresponding screw hole. Existing vision or sensor-based object tracking algorithms commonly fail when parts are occluded or have a small footprint [67]. To address this technical issue, EditAR allows authors to embed 2D video recordings as an alternative to 3D instructions. Through the self-determined hand gestures, the author controls the beginning and end of a video. EditAR captures and stores real-time video through the ZED camera attached to the HMD. In VR and AR, the video is then embedded in 3D space based on the user's current head pose. In the video, these recordings are added to the video clip.

4.2.3 Step 3: Create 2D video from Digital Twin

For creating a 2D instructional video, the user uses the digital twin that was created in the previous step. For this, EditAR provides the user with a comprehensive, immersive 3D video interface to edit and produce 2D video renderings from the Digital Twin.

3D Editor: The EditAR interface includes several standard features similar to a desktop-based video editor directly in the virtual environment, including loading recordings, scrubbing through the timeline, and previewing. In addition to these features, the editor supports novel 3D interactions that are useful for creating video content from a 3D recording. These include several virtual camera manipulations and the ability to hide the editor, walk around the scene, and show the preview screen at different locations. Combined with the traditional video features, the new 3D video editing features allow EditAR to offer familiar video editing and to enable video content production from pre-recorded virtual content in a novel way.

Virtual cameras: The core of EditAR's 3D video editor is the virtual camera that allows for the 3D recordings to be captured as 2D video. Virtual cameras are common in VFX, movie-making, and gaming, but their interfaces are designed for expert users using a desktop. To adapt this technology for in-headset free-hand interaction editing, we performed a pilot study to inform our final design. Based on broad categories of input given by multimedia experts and our survey of various kinds of physical camera metaphors already in existence, we developed ten virtual camera variations[1], illustrated in Fig. 4 and described briefly in Table 1.

The static third-person camera and manual camera (Table 1) functions are supported on Unity 3D natively. Other manual cameras are inspired by real-world cameras or rigs but are algorithmically supported. The manual track camera transforms a virtual joystick movement, where a straight and arc line follows a predetermined path set by the user. For the moving focal point, the user determines the trajectory of the focal point instead of the camera path trajectory. Automated hands[2] use the position of the hands to remain in focus.

[1]Working footage of these cameras are provided in the supplementary material to help the reader visualize the functions

[2]Implementation Algorithm provided in supplementary material

		Camera Type	Description	Ranking
Static		Third Person	Static camera to provide a single third-person perspective (Fig. 4 (D))	2
		CCTV	**Multiple Static Camera angles that the user can switch between during recording (Fig. 4 (E))**	**1**
Dynamic	Automatic	**First Person**	**As seen from the eyes of the author (Fig. 4 (B))**	**1**
		Automated Hand Tracking	The camera moves automatically to always keep both hands within the frame (Fig. 4 (G))	2
	Manual	Manual	Manually manipulate camera with hands (Fig. 4 (A)	4
		Interpolation 3rd person	**The camera sweeps between multiple manually set camera positions during recording (Fig. 4 (F))**	**1**
		Manual Track	Manually manipulate the camera with a joystick (Fig. 4 (C))	5
		Straight Line	Camera travels along a straight line created by user (Fig. 4 (H))	6
		Arc Line	The c travels along an arc about a fixed focal point with radius set by the user (Fig. 4 (I))	3
		Moving Focal point	Camera pivots about a point as the focal point moves along a line created by the user (Fig. 4 (J))	2

Table 1: Cameras tested during the preliminary study.

Pilot User Study: The goal of this preliminary study was to evaluate the set of 10 virtual cameras that we had developed. We invited ten users (all male) for a 90-minute session; each was compensated with a USD 15 e-gift card. The participants had varying degrees of familiarity with video filming (one user who edits and films videos almost daily, two who film a few times a week, six who reported filming a few times a month, and one user with no experience).

After a brief explanation of the study, participants signed a consent form. Next, they provided demographic information in a questionnaire. No user had previous interaction experience in VR. Then, the experimenter demonstrated how to use the virtual cameras and their functions. Next, the users were given time to practice with all the virtual cameras to record 2D videos of a pre-recorded 3D recording (Digital twin), illustrating a tee-joint welding action. We recorded the learning time for each camera and the number of practice attempts made before users were satisfied with the video.

After practice, each participant was provided with a pre-recorded 3D scene of a bike repair operation, and asked to capture a 2D video of the 3D recording with each virtual camera. The task completion time and the number of attempts with each camera to capture a satisfactory video were again recorded. After the use of each of the cameras, the users filled a 7-point Likert scale questionnaire to evaluate the camera's usability and their satisfaction. After using all ten virtual cameras, the users ranked each of the cameras based on the quality of the final videos, ease of use, and usefulness. An interview with open-ended questions followed this.

The ordinal ranking data were pooled for all users and compared against each other, to identify the best-suited camera for each category, see Table 1. The best-ranked camera from each category (Static, Dynamic-automated, and Dynamic-manual) was selected for inclusion in the final version of EditAR. The information obtained from this pilot study could also inform further research on virtual cameras controlled with free-hand interaction in immersive systems.

The results of the pilot user study encouraged us to include three cameras for recording the content in the 3D Editor: First-Person, Interpolation, and CCTV. The First-Person camera effectively aligns the camera with the author's head position, providing a first-person view of the task. The Interpolation camera allows the user to place a camera anywhere in the scene at any given time, as identified by scrubbing the timeline. The editor will then interpolate the camera pose between the specified positions, creating a cinematic feel to the camera rendering. The last camera type is CCTV, which allows the user to place four separate cameras in the scene at desired positions. The recording slice is then played back using a four-camera view (similar to a CCTV display), and the user can select which camera to "record" out of the four previews available (similar to television production for sports [44]). This allows the user to see several different points of view for the same task, allowing them to choose the best point of view at the current time.

Throughout the entire video editing process, a preview of the final 2D video can be seen on a small display in the middle of the editor, much like a traditional video editor. As the user scrubs through the timeline, the scene will update, and they can see where exactly their virtual camera is located at that instant. Lastly, once the user completes their video editing process, they can play their final product back and watch it on a standard monitor and export it for playback on other devices.

5 EditAR Evaluation

5.1 Expert Review

To assess the overall system and effectiveness of the system, we evaluated EditAR with 16 *Expert Users* with varying multimedia and subject expertise levels. Because EditAR caters to different multimedia users, we recruited four experts for each media type (AR, VR, and video), and four subject experts with expertise in welding (a spatial skill).

Study setup: The participants (15 male; 1 female, ranging in age from 21 to 35 years old), received USD 20 for the 2-hour session.

After an initial demographics survey, the participants were split into four user groups. The first three user groups were tasked with authoring an instruction set for assembling an aluminium 80/20 rig involving eight sub-tasks. For reference, a printed instruction manual for this task was provided. All participants in the first three user groups were given a 15-minute assembly tutorial and asked to practice the task. The fourth user group consisted of subject experts with expertise in welding, and we consequently used a welding scenario for this group.

Subsequently, users were asked to create instructions for the task they had just learned. The chosen assembly task contains elements of kinesthetic learning, such as interacting with tools and performing assembly operations, and we made sure that it is simple enough for non-subject experts (but still multimedia experts) to learn.

Measures: As there are currently no standard tools for authoring content for three multimedia types simultaneously, we did not perform a direct one-on-one comparative study. As the user study already involved a 2-hour session, it was also impractical to use additional time to teach other authoring modes to users without the corresponding expertise. Hence, we used different media types as control conditions for different groups. The last group of users with no media expertise only performed the task with EditAR. We still decided to have the participants in the first three groups create instructions as a control condition to obtain a baseline for the amount of time required for creating instructions in traditional media. We measured task success and task completion time (including each sub-task completion).

Overall, the first three user groups were experts in their respective multimedia fields, and all had prior experience with their respective baseline conditions. Because we were interested in obtaining qualitative feedback only for EditAR, the NASA TLX [15], SUS [5] were administered only for the experimental condition. We ended the study with an unstructured interview to gather further insights. All actions were captured with video cameras, and the user's point of view was recorded with a screen capture tool. Interviews were recorded for later analysis.

User Group 1: Group 1 consisted of four AR experts, all of whom had experience with developing phone-based AR content with the ARcore (3 users) and ARKit (1 user) platforms. The users were asked to create instructions in two conditions.

Condition 1: Here, participants were asked to create AR content using a provided AR authoring system, which we developed to be equivalent to commercially available systems, such as PTC Vuforia capture [45] and Microsoft Dynamics 365 [31]. With this approach, users have to first capture the physical task using a camera (Pixel 4 and tripod provided). Then, they have to transfer the video into

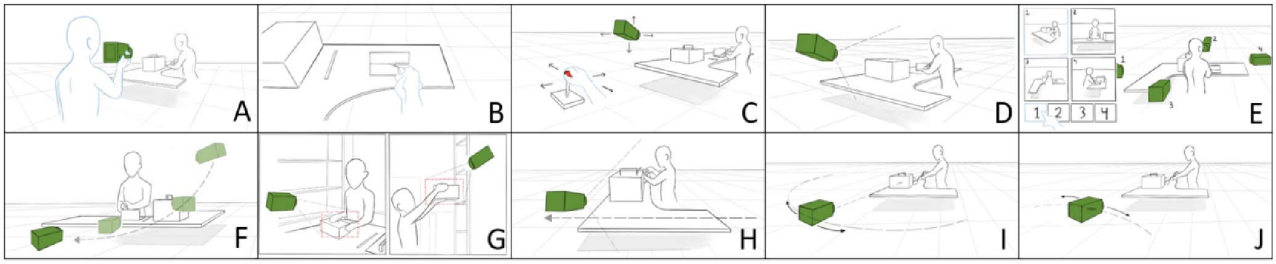

Figure 4: Ten cameras tested during the preliminary study. Manual (A), First Person (B), Manual Track (C), Third Person (D), CCTV (E), Interpolation 3rd Person (F), Automate Hand Tracking (G), Straight Line (H), Arc Line (I), and Moving Focal Point (J).

User Group	Avg. Time (mins)	Avg. Raw NASA TLX	Avg. SUS
1 - EditAR	24	34.6	79.9
1 - AR	21		
2 - EditAR	25	34.7	86.7
2 - VR	N/A		
3 - EditAR	23	37	78.4
3 - Video	24		
4 - EditAR	18	33.5	79.4

Table 2: Average Raw NASA TLX and SUS scores for all user groups for the EditAR condition along with average task completion times for all conditions.

Unity via the desktop interface. Finally, the users need to embed the instruction video into the real world in the AR view. Before the task, users were given 10 minutes to familiarize themselves with the system. After that, the AR experts were asked to author only AR instructions under this condition.

Condition 2: The experimenter first introduced the EditAR environment by demonstration. The users were then given time to practice with the different features until they were comfortable using the system without assistance. Then, the users were asked to author instructions for all three media with EditAR.

User Group 2: Group 2 consisted of four VR experts recruited from a local VR club that focuses on developing VR games. All users reported a minimum of 12 months of experience with Unity and the Oculus headsets. Group 2 experienced two conditions. In condition 1 they were asked to develop VR instruction for the assembly task. We decided to perform a comparison with Unity 3D, as most current state-of-the-art commercial VR authoring tools, such as Tvori [60] and TiltBrush [12], do not support real-world interaction when developing virtual content, and this disconnect discouraged us from using them as a baseline. For all tools and objects in the environment, virtual models were provided. Condition 2 was the same as Group 1's condition 2, and thus the VR experts were asked to author instructions for all three media with EditAR.

User Group 3: User group 3 consisted of 4 videographers (1 travel vlogger, two educational video developers, and one part-time wedding videographer). All users reported familiarity with video editing software, Camtasia or Davinci Resolve 17, and a minimum of 24 months of video editing experience. Group 3 experienced the following two conditions. In condition 1 the users were asked to develop video instructions for the assembly task with video editing software (Camtasia or Davinci resolve). Condition 2 was the same as the previous group's condition 2, and thus the video experts were asked to author instructions for all three media with EditAR.

User Group 4: Group 4 contained 4 SMEs with expertise in welding, with expertise ranging from 15 months to 6 years. One user reported owning an Oculus Quest 1 VR headset and using it for playing VR games with controllers. Due to their expertise, these participants were asked to create content for all three media using EditAR as the only condition. Yet, unlike the previous groups and the assembly use case, these experts were asked to create content for a welding task, specifically for metal feed welding two steel flats. Due to laboratory fire safety protocols, the experts were asked not to actually weld but to simulate the welding action to create content.

Results: The average task completion times and raw NASA TLX and SUS scores for the EditAR conditions are shown in Table 2. All users from Groups 1, 3, and 4 could complete all sub-tasks. However, all four VR experts ran out of time when generating VR instructions for all eight sub-tasks with the conventional approach. As we needed to fit the whole study into the allotted 2 hours, the experimenter asked participants to stop authoring VR after 40 min had elapsed. One user completed 7 sub-tasks, two users completed 6 sub-tasks, and one user completed 5 sub-tasks. Yet, all four users finished authoring all three media versions in the EditAR condition.

5.2 Discussion

Overall, the task success rate and completion times for the SMEs (user group 4) is promising evidence for our claim that EditAR allows users with no formal multimedia expertise to quickly create content for AR, VR, and video media, based on only a single demonstration. In addition, the average SUS score for EditAR in user group 4 was 82. This is encouraging, as an average score of 70 and above translates to "excellent" usability [3].

A statistical analysis of the average time for authoring instruction between EditAR and traditional video editing did not identify a significant difference. Because the users can render content for *three different media* with EditAR within (almost) the same amount of time as for a *single* traditional medium, this might not be a fair comparison. We addressed this issue in our next study (see Sect. 6). Participants appreciated the novelty of EditAR and that experts and novices (in terms of multimedia creation) alike could successfully create multimedia content with little training. That said, experts who had ample experience with multimedia tools still offered exciting comments. For example, the availability of different camera perspectives was validated, as SMEs and videographers both identified the difficulty of showing scenes from multiple perspectives–especially to effectuate corrections and re-shoots. As highlighted by multimedia experts after the user studies, the ability to go back and shoot part of the recording from a different perspective and then incorporate that into the original project simplified the entire creation process in a highly appreciated way. The camera perspectives provided by EditAR–based on our formative interviews with video experts–were also commented on positively by videographers.

An interesting observation from the study was that none of the VR experts could complete the authoring of all sub-tasks for assembly instruction in the Unity environment with which they were intricately familiar with. Comments from the post-study interview provided some interesting insights here. E5 said, *"yeah, [EditAR] was faster for what I was trying to do. I was trying to program [a] tool path like slide the 80/20 or rotate the screw driver... which was hard to do within the given time"*. E7 stated, *"Programming precise*

operation[s] is hard, so I guess directly showing how to do [it] is a clever solution." Overall, we found that EditAR enables non-multimedia experts to author instructions for kinesthetic tasks with just a few minutes of training.

6 Comparative Evaluation: EditAR vs. Traditional Video Creation

Participants We invited eight participants (U1-U8) (two female, six male) for a two-session study lasting approximately 90 minutes each. They were compensated with USD 15 e-gift cards for each session. The users were 18-34 years old. All users had no prior experience with video editing. Two users reported using VR headsets to play games with controllers 1-2 times in the past year. However, they reported no freehand interaction experience. All users reported familiarity with using a keyboard and mouse interface, using them almost every day in the past year.

Study Design We designed a counterbalanced, within-subjects study where users experience two conditions: EditAR and traditional video creation. They were first trained and then asked to create an instruction video of the assembly task under both conditions. We chose an assembly task of an 80/20 aluminum frame which requires five specific sub-tasks. This task is both representative of a typical assembly operation in an DIY project and also of activities using kinesthetic instructions.

Procedure The study was conducted in the lab. A brief explanation was provided to the participants followed by consent form and a demographic questionnaire. For EditAR, the researcher demonstrated each feature. Then, participants were asked to practice on a pre-recorded Digital Twin. Up to 20 minutes were allocated for training, with individual times shown in Figure 5. The number of accidental gesture inputs by participants was also recorded. For the video condition, which used DaVinci Resolve 17 [4], the experimenter provided a tutorial on basic video editing operations such as loading, splitting, deleting, and repositioning video clips on a traditional WIMP [61] environment. Similar to the other condition, up to 20 minutes of training were given. The users were provided with sample footage for practice.

Subsequently, the users were taught the assembly operation and asked to practice it. In the EditAR condition, the users were provided with Antilatency sensors for their 3D recording. They were then asked to preform the task to generate the digital twin. In the EditAR conditions, participants were asked to use EditAR's 3D editor and to create a 2D video from their 3D recordings. In the video condition, the users were provided with a tripod and a Google Pixel 3 to capture video footage. The number of physical camera movements was recorded during the filming process. Task completion time was measured from when the participant said they were ready until they were satisfied they had completed each 3D/2D recording, inclusive editing.

Measures After each session, we administered a SUS survey [5] to measure usability and a NASA-TLX [15] for perceived workload. Then, a 7-point Likert scale questionnaire was used to gather qualitative feedback on the system. Finally, to elicit their perceptions, we performed an open-ended interview.

6.1 Results and Discussion

We collected the amount of time taken for learning, practice, 3D recording, editing, and total task completion, alongside usability and cognitive load ratings for both conditions (EditAR and Video) and analyzed this data using paired sample t-tests (refer to Figure 6). For all statistical tests, $p<.05$ was considered significant.

Although the average total task completion time with EditAR (M=20.3 mins, SD=10.49) is numerically lower than with traditional video (M=32.64 mins, SD=25.22), the difference was not statistically significant, t(7)=2.14, p=.069. The amount of time taken to edit in both conditions was similar (EditAR: M=17.65 mins, SD=9.57, and video: M=17.38 mins, SD=14.27) and not statistically significance, t(7)=0.10, p=.92. However, the 3D recording time exhibited a significant difference, t(7)=3.64, p=.0083 between EditAR (M=2.66 mins, SD=1.32) and traditional video (M=15.25 mins, SD=11.01), with EditAR able to capture content faster. Finally, the amount of time each user spent to learn the system also was significantly different, with traditional video (M=1.89 mins, SD=1.42) being significantly faster than EditAR (M=14.52 mins, SD=3.09), with t(7)=12.99, p=.0001.

The SUS usability scores of EditAR (M=77.5, SD=13.09) were significantly higher than traditional video (M=54.6, SD=20.37) t(7)= 2.88, p=.02. However, there was no significant difference in terms of the NASA TLX (EditAR: M=59, SD=14.62; Video: M=59.5, SD=8.67), t(7)=0.1, p=.92.

Workflow Efficiency: Our primary research goal for EditAR was to develop an efficient and unifying instruction authoring workflow. The fact that total task completion time and the editing time showed no significant difference between EditAR and traditional video conditions supports our primary objective of developing an efficient workflow. As EditAR can create content for *three different media simultaneously* in the same time it takes to authoring a *single* traditional video, this validates our contribution.

For traditional video, three actions are required to successfully capture an instructional set: Performing the task, filming, and editing. As the "physical task performance" is captured as a digital recording in EditAR, which enables reuse, this increases workflow efficiency. In addition, digital content enables authors to refer back to their 3D actions at any time. Combined with EditAR's 3D video editor, which permits "filming" and "editing" to be performed together in the same system instead of in-sequence in different systems, this also makes the workflow more efficient.

The following participant comments best illustrate the strength of EditAR. U4 stated: *"when I did the video recording thing, there were many instances when I forgot to start recording, or I was pointing it in the wrong direction. And I had to distribute my attention [] between actually doing the task with my hands, and [] also stretching my neck and looking into the screen, if it's [] right, it's focused in the correct place on you know, if my hands and my actions are even captured by the screen, so such things will definitely not occur with editor because you always have this birds eye view and the small screen, the preview screen, where you can see whichever angles you want to see. So yeah, that was pretty cool."* U3 elaborated: *"In VR [i.e., EditAR] placing cameras and switching between different types of cameras is new, that is much more easier. Because in video filming, we have to place a lot of cameras, and then it takes up a lot of time. So sometimes people tend to miss out on a few angles."* U3's comment is further corroborated by a comparison of the number of times the physical camera was moved during the traditional video condition (M=7.38, SD=2.37) and the number of times different virtual cameras were used in EditAR (M=5.88, SD=2.64). This difference is statically significant, t(7)=5.61, p=.0008, with higher demand for the video condition.

Usability and Cognitive Load: Based on the SUS scores, the participants clearly felt EditAR to be more usable than the traditional video interface. However, there was no significant difference for self-reported cognitive load as measured by the NASA TLX. These results corroborate the findings of Griffin et al. for 6Dive [13], in terms of the perceived cognitive load for in-headset based editing using 6DoF controllers. EditAR still extends their work in terms of the usability of the system, as freehand interaction and gestures improved the usability of an in-headset based editing interface over the reliance on 6DoF controllers. Here it is interesting to point out that the *subject matter experts* from the ProcessAR [7] study identified the use of controllers as a limitation. Our study provides evidence that freehand interaction improves the usability of AR instruction set authoring.

	EditAR					Video		
User No.	3D Recording Time (mins)	Editing Time (mins)	Practice Time (mins)	Accidental Input (Practice)	Accidental Input (Study)	2D Recording Time (mins)	Editing Time (mins)	Practise Time (mins)
1	1.32	7.27	11.4	8	2	3.44	3.48	0.75
2	1.75	11.65	10.73	3	1	6.94	7.97	1.11
3	2.13	32.3	19.32	5	2	19.33	23.9	1.98
4	4.55	22.88	16.65	4	1	26.21	27.83	1.55
5	4.9	30.55	16.97	3	2	35.63	45.95	5.25
6	2.28	13.5	15.76	4	3	10.23	9.51	1.52
7	2.43	11.7	12.07	4	2	12.52	12.58	1.9
8	1.9	11.45	13.22	4	1	7.72	7.88	1.05

Figure 5: Amount of time spent by users for practice, recording, and editing for both EditAR and Video conditions, along with the number of accidental gesture inputs during practise and study sessions for the EditAR condition.

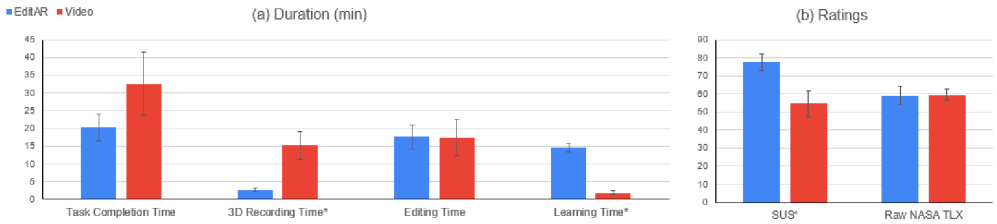

Figure 6: Comparison between the EditAR and Video conditions. (a) Amount of time spent on different tasks during content creation. (b) Average Usability and Cognitive load ratings. Significant results marked with *.

Learnability and Accidental Inputs: While EditAR has its strengths, there are areas where traditional video systems still prevail. One is learnability: users took significantly more time to learn the EditAR interface compared to traditional video. We attribute this to the universally high familiarity with a keyboard and mouse interface. A related issue is the somewhat higher error rate due to accidental gesture detection in EditAR. When using EditAR, both during practice and editing, users sometimes performed unintended or accidental hand movements or gestures that were recognized as input. The number of such observations is reported in Figure 5. As the users became more accustomed to the interface over time, we noticed that the number of such observations dropped, from M=4.38, SD=1.60 during the practice session to M=1.75, SD=0.7 during editing, which means that this might not be "just" a recognition issue. We predict that as AR/VR hand-based interaction and gestures become more commonplace in the future, we can expect improvements in terms of accidental inputs and learning time. Yet, further research is required to verify this.

7 Limitations, Future Work and Conclusion

EditAR also presents a baseline for future research in multiple-media authoring. For example, while maintaining EditAR's initial 3D capture workflow, a comparative evaluation for the subsequent video clip editing between desktop, WIMP-based interfaces, and in-headset-based interfaces might be valuable. Such a study would be informative even beyond the scope of multiple-media authoring, and thus requires more careful evaluation and analysis.

We currently use 6DoF sensors, which afford reliable 3D pose tracking to provide high fidelity digital capture and to avoid missing point cloud information due to occlusion. Yet, with advancements in computer vision based on 3D cameras, such as explored in Fender et al. [9], such technology could accomplish the same task without the aid of external sensors. Still, hand occlusions (e.g., for the tool held in a hand) are known to affect such approaches. Also, substantial data, time, and effort is required to train such systems for highly reliable tracking. To avoid problems due to all these issues, we did not use such an approach in our user studies.

We decided to use a single camera setup for our comparative study because this option is frequently used in movie [23] and video content creation (e.g., for YouTube). A single camera also reduces user effort, as planning multiple camera angles takes expertise and additional time [35], and minimizes complexity in controlling lighting and shadows [32]. EditAR enables video creation from the captured digital twin, eliminating challenges due to limitations associated with physical light sources and permitting the user to optimize camera angles after the fact. Still, it might be interesting to perform a comparison with a multi-camera setup in future work. The results presented here will still act as a baseline for further research in this space.

We presented EditAR, an AR-based editing platform for authoring asynchronous kinesthetic instructions in multiple media formats. Our user evaluations with both multimedia experts, SMEs, and novices demonstrate that creators with no experience in AR or VR content creation or filming videos can easily use our system to author instructions for multiple media formats with only minimal training. EditAR thus enables more SMEs to create instructional content and share their hands-on skills and knowledge. The generated content is then accessible asynchronously to a broader audience, reducing costs and making training scalable across various tasks and procedures. We thus provide an innovative solution to the problem of re-skilling and up-skilling the future workforce and hope that our findings inspire future work in this area.

Acknowledgments

We thank Quentin LaFollette and David L Chapman for their help with the sketches. This work was partially supported by the NSF under grants Future of Work at the Human Technology Frontier (FW-HTF) 1839971. We also acknowledge the Feddersen Distinguished Professor Funds. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] F. Anderson, T. Grossman, J. Matejka, and G. Fitzmaurice. Youmove: enhancing movement training with an augmented reality mirror. In *Proceedings of the 26th annual ACM symposium on User interface*

software and technology, pp. 311–320. Association for Computing Machinery, New York, NY, USA, 2013.
[2] Antilatency. Antilatency, Apr. 2021.
[3] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.
[4] Blackmagicdesign. Davinci resolve 17, Nov. 2020.
[5] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
[6] Y. Cao, X. Qian, T. Wang, R. Lee, K. Huo, and K. Ramani. An exploratory study of augmented reality presence for tutoring machine tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. Association for Computing Machinery, New York, NY, USA, 2020.
[7] S. Chidambaram, H. Huang, F. He, X. Qian, A. M. Villanueva, T. S. Redick, W. Stuerzlinger, and K. Ramani. Processar: An augmented reality-based tool to create in-situ procedural 2d/3d ar instructions. In *Designing Interactive Systems Conference 2021*, DIS '21, p. 234–249. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3461778.3462126
[8] L. L. Cone. Skycam: an aerial robotic camera system. *Byte Magazine*, 10:122, 1985.
[9] A. R. Fender and C. Holz. Causality-preserving asynchronous reality. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3501836
[10] M. Funk, M. Kritzler, and F. Michahelles. Holocollab: a shared virtual platform for physical assembly training using spatially-aware head-mounted displays. In *Proceedings of the Seventh International Conference on the Internet of Things*, pp. 1–7. Association for Computing Machinery, New York, NY, USA, 2017.
[11] Q. Galvane, I.-S. Lin, F. Argelaguet, T.-Y. Li, and M. Christie. Vr as a content creation tool for movie previsualisation. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 303–311, 2019. doi: 10.1109/VR.2019.8798181
[12] Google. Tilt brush, Apr. 2016.
[13] R. Griffin, T. Langlotz, and S. Zollmann. 6dive: 6 degrees-of-freedom immersive video editor. *Frontiers in Virtual Reality*, 2:75, 2021.
[14] T. Ha, W. Woo, Y. Lee, J. Lee, J. Ryu, H. Choi, and K. Lee. Artalet: tangible user interface based immersive augmented reality authoring tool for digilog book. In *2010 International Symposium on Ubiquitous Virtual Reality*, pp. 40–43. IEEE Computer Society, Los Alamitos, CA, USA, 2010.
[15] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, USA, 1988.
[16] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 191–200. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 2011.
[17] K. Huang, J. Li, M. Sousa, and T. Grossman. Immersivepov: Filming how-to videos with a head-mounted 360° action camera. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517468
[18] A. Hughes. Forging the digital twin in discrete manufacturing: A vision for unity in the virtual and real worlds, September 2018.
[19] T. Imai, A. E. Johnson, J. Leigh, D. E. Pape, and T. A. DeFanti. The virtual mail system. In *Proceedings of Virtual Reality*, pp. 78–78. IEEE Computer Society, Los Alamitos, CA, USA, 1999.
[20] instructables. instructables, Apr. 2021.
[21] A. Ipsita, L. Erickson, Y. Dong, J. Huang, A. K. Bushinski, S. Saradhi, A. M. Villanueva, K. A. Peppler, T. S. Redick, and K. Ramani. Towards modeling of virtual reality welding simulators to promote accessible and scalable training. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517696
[22] S. Kim, H. gun Chi, and K. Ramani. Object synthesis by learning part geometry with surface and volumetric representations. *Computer-Aided Design*, 130:102932, 2021. doi: 10.1016/j.cad.2020.102932
[23] P. Kiwitt. What is cinema in a digital age? divergent definitions from a production perspective. *Journal of Film and Video*, 64(4):3–22, 2012.
[24] L. Lee, T. Braud, P. Zhou, L. Wang, D. Xu, Z. Lin, A. Kumar, C. Bermejo, and P. Hui. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *CoRR*, arXiv:2110.05352, 2021.
[25] G. Leiva, C. Nguyen, R. H. Kazi, and P. Asente. Pronto: Rapid augmented reality video prototyping using sketches and enaction. CHI '20, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376160
[26] K. Lilija, H. Pohl, and K. Hornbæk. Who put that there? temporal navigation of spatial recordings by direct manipulation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–11. Association for Computing Machinery, New York, NY, USA, 2020.
[27] S.-C. Lim, H.-K. Lee, and J. Park. Role of combined tactile and kinesthetic feedback in minimally invasive surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 11(3):360–374, 2015.
[28] S. Mallery. Glidetrack camera slider rigs, 2011. https://www.bhphotovideo.com/explora/photography/hands-review/glidetrack-camera-slider-rigs.
[29] M. R. Marner, A. Irlitti, and B. H. Thomas. Improving procedural task performance with augmented reality annotations. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 39–48. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 2013.
[30] A. Marwanto, Y. Wibowo, R. Djatmiko, and R. Wijaya. Kinesthetic intelligence in welding practice lectures. In *Journal of Physics: Conference Series*, number 1, p. 012022. IOP Publishing, IOP Publishing, Orlando, Florida, 2020.
[31] Microsoft. Overview of dynamics 365 guides, 2020. Retrieved May 5,2020, from `https://docs.microsoft.com/en-us/dynamics365/mixed-reality/guides/`.
[32] G. Millerson. *Lighting for TV and Film*. Routledge, 2013.
[33] P. Mohr, D. Mandl, M. Tatzgern, E. Veas, D. Schmalstieg, and D. Kalkofen. Retargeting video tutorials showing tools with surface contact to augmented reality. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 6547–6558. Association for Computing Machinery, New York, NY, USA, 2017.
[34] M. Morozov, A. Gerasimov, and M. Fominykh. vacademia–educational virtual world with 3d recording. In *2012 International Conference on Cyberworlds*, pp. 199–206. IEEE Computer Society, Los Alamitos, CA, USA, 2012.
[35] R. B. Musburger. *Single-camera video production*. Routledge, 2012.
[36] M. Nebeling, S. Rajaram, L. Wu, Y. Cheng, and J. Herskovitz. Xrstudio: A virtual production and live streaming system for immersive instructional experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3411764.3445323
[37] C. Nguyen, S. DiVerdi, A. Hertzmann, and F. Liu. Vremiere: In-headset virtual reality video editing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, p. 5428–5438. Association for Computing Machinery, New York, NY, USA, 2017.
[38] nvrmind. nvrmind, Feb. 2022.
[39] Oculus. Oculus software development kit, 2019. Retrieved September 18, 2019, from `https://developer.oculus.com`.
[40] Oculus. Oculus quest 2, 2020. Retrieved April 4, 2021, from `https://www.oculus.com/quest-2/`.
[41] O. Oda, C. Elvezio, M. Sukan, S. Feiner, and B. Tversky. Virtual replicas for remote assistance in virtual and augmented reality. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 405–415. Association for Computing Machinery, New York, NY, USA, 2015.
[42] S. Ong and Z. Wang. Augmented assembly technologies based on 3d bare-hand interaction. *CIRP annals*, 60(1):1–4, 2011.
[43] OptiTrack. Optitrack, Apr. 2021.
[44] M. Perry, O. Juhlin, M. Esbjörnsson, and A. Engström. *Lean Collaboration through Video Gestures: Co-Ordinating the Production of Live*

Televised Sport, p. 2279–2288. Association for Computing Machinery, New York, NY, USA, 2009.
[45] PTC. Vuforia expert capture, 2019. Retrieved May 5,2020, from https://www.ptc.com/en/products/augmented-reality/vuforia-expert-capture.
[46] PTC. Vuforia chalk: Remote assistance powered by augmented reality, Dec. 2020.
[47] quillart. quillart, Feb. 2022.
[48] R. Radkowski and C. Stritzke. Interactive hand gesture-based assembly for augmented reality applications. In *Proceedings of the 2012 International Conference on Advances in Computer-Human Interactions*, pp. 303–308. Citeseer, Valencia, Spain, 2012.
[49] M. Sayed, R. Cinca, E. Costanza, and G. Brostow. Lookout! interactive camera gimbal controller for filming long takes, 2020.
[50] M. Scheff-King. Download, edit and print your own parts from mcmaster-carr, Apr. 2014.
[51] M. Sheets-Johnstone. Kinesthetic memory. *Theoria et historia scientiarum*, 7(1):69–92, 2003.
[52] Skillshare. Skillshare, Apr. 2021.
[53] Stereolabs. Zed mini, Apr. 2021.
[54] stratasys. Grabcad community, 2022. Retrieved March 8, 2022, from https://www.traceparts.com/en.
[55] R. Suzuki, R. H. Kazi, L.-y. Wei, S. DiVerdi, W. Li, and D. Leithinger. Realitysketch: Embedding responsive graphics and visualizations in ar through dynamic sketching. UIST '20, p. 166–181. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3379337.3415892
[56] TeamViewer. Teamviewer pilot, Dec. 2020.
[57] B. Thoravi Kumaravel, F. Anderson, G. Fitzmaurice, B. Hartmann, and T. Grossman. Loki: Facilitating remote instruction of physical tasks using bi-directional mixed-reality telepresence. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 161–174. Association for Computing Machinery, New York, NY, USA, 2019.
[58] Toribash. Toribash, Mar. 2006.
[59] traceparts. Traceparts, 1990. Retrieved March 8, 2022, from https://www.traceparts.com/en.
[60] tvori. tvori, Feb. 2022.
[61] A. van Dam. Post-wimp user interfaces. *Commun. ACM*, 40(2):63–67, feb 1997. doi: 10.1145/253671.253708
[62] vreal. vreal, Mar. 2018.
[63] C. Y. Wang, M. Sakashita, U. Ehsan, J. Li, and A. S. Won. Again, together: Socially reliving virtual reality experiences when separated. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. Association for Computing Machinery, New York, NY, USA, 2020.
[64] T. Wang, X. Qian, F. He, X. Hu, Y. Cao, and K. Ramani. *GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications*, p. 552–567. Association for Computing Machinery, New York, NY, USA, 2021.
[65] M. Whitlock, G. Fitzmaurice, T. Grossman, and J. Matejka. Authar: Concurrent authoring of tutorials for ar assembly guidance. In *Graphics Interface*, pp. 431 – 439. CHCCS/SCDHM, University of Toronto,Ontario, Canada, 2020.
[66] L. Wright and S. Davidson. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, 7(1):1–13, 2020.
[67] M. Yamaguchi, S. Mori, P. Mohr, M. Tatzgern, A. Stanescu, H. Saito, and D. Kalkofen. Video-annotated augmented reality assembly tutorials. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 1010–1022. Association for Computing Machinery, New York, NY, USA, 2020.
[68] U. Yang and G. J. Kim. Implementation and evaluation of "just follow me": An immersive, vr-based, motion-training system. *Presence*, 11(3):304–323, 2002. doi: 10.1162/105474602317473240
[69] YouTube. Youtube, Apr. 2021.
[70] J. Zillner, E. Mendez, and D. Wagner. Augmented reality remote collaboration with dense reconstruction. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 38–39, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00028