Title	Bridged reality : a toolkit for live holographic point cloud data interaction
Sub Title	
Author	Armstrong, Mark(Minamizawa, Kōta) 南澤, 孝太
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2021
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2021年度メディアデザイン学 第865号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002021-0865

Master's Thesis
Academic Year 2021

Bridged Reality: A Toolkit for Live Holographic Point Cloud Data Interaction

Keio University
Graduate School of Media Design

Mark Armstrong

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Mark Armstrong

Master's Thesis Advisory Committee:

Professor Kouta Minamizawa	(Main Research Supervisor)
Professor Kai Kunze	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Kouta Minamizawa	(Chair)
Professor Kai Kunze	(Co-Reviewer)
Professor Hideki Sunahara	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2021

Bridged Reality: A Toolkit for Live Holographic Point Cloud Data Interaction

Category: Science / Engineering

Summary

In the current evolution of augmented reality, there is an emphasis on use in entertainment, autonomous vehicles, and information displays. There are a surplus of environments that could benefit from digital augmentation. Most commercial AR systems are designed for an egocentric HMD experience, and lack compatibility for custom virtual controller interaction. Through the emergence and accessibility of RGB-D cameras, there are an increasing number of opportunities for humans to integrate the physical world into the virtual, and to reflect the virtual world back into an augmented physical space.

Bridged Reality is a toolkit that aims to popularize cross-reality environments and foster uncommon depth-based interactions that result in enchanting virtual effects. The toolkit tracks user data in an augmented environment, rapidly processes the live data, and displays the output across multiple holographic screens. This work presents a localized graphical rendering technique, as well as a closed input driven feedback loop between virtual and physical environments. The application of this work could lead to users viewing and interacting with their live selves in video games, and interactive holographic exhibitions.

Keywords:

hybrid-spaces, point cloud interaction, virtual effects, holographic displays, virtual controllers

Keio University Graduate School of Media Design

Mark Armstrong

Contents

List of Figures

Acknowledgements

Most importantly, thank you Yusa for being my rock and helping me through this very difficult time in my life with your infinite kindness, enthusiasm, and loving energy.

Thank you to those who raised me from my roots - my late parents, Barbara, Brianna, LJLC, Kerns, Parkers. And to Jeff for helping me realize one of my life's goals: coming to Japan.

Thanks to my KMD faculty for babysitting me through this Master's Program, Pai, Yamen, Kai, and Minamizawa. Thank you Anish for helping me lay out this thesis when I had absolutely no clue what I was doing. Thanks #m2019f for your cooperation.

Shoutout to KFC, TCC, my dog Lea, Pilam, PALS, Captain Jack Sparrow, Elon Musk, Emotional Oranges, and all the other cool and inspirational people I sometimes look up to that I'm certain will never read this.

Chapter 1
Introduction

1.1. How Humans Interact with a Virtual World

Never before in the history of humanity has it been as desirable and fulfilling as it is today to stay at home, locked in your room, disconnected from the rest of the physical world. The obvious enabler of this lifestyle: computers. While a stable internet connection serves as a link to remain connected with all of our social groups, games and applications that don't require an internet connection are also sufficient as activities to combat stress, to work, to create art, and to share cultural values. But why is it more preferable to stay at home for some then to go outside and experience the physical world?

Different reasons for different people - there is no right answer. For some it may be that they can behave more naturally and instinctively at home without societal pressure. For others it may be their disposition to enjoy existing in a familiar environment. For many, being housebound is not a choice but a result of unfortunate circumstances. For this same group, their familiar environment may be small, unengaging, and boring. What options do they have to make their base of operations more desireable and enjoyable? The introduction of virtual reality systems nearly 60 years ago made it attainable from a visual perspective to escape physical reality entirely, and to exist in a more desirable space [1]. This opportunity was delivered regardless of a user's personal preference towards their environment.

But viewing another space, and actually existing and operating in it are two completely different checkpoints. A significant part of our interaction with computers is our comfortability and familiarity with our controllers. The different controllers that we use will change our perception of our human-computer experiences. Controllers are generally classified based on their shape and use cases.

PlayStation controllers [1] are typically used for casual console gaming and simple robotics, while keyboards are more often used in contexts like competitive gaming and work. Non-conventional controllers lay to foundation to non-conventional digital experiences.

Types of controllers

A controller is simply an interface through which a user can cause an expected output. Just as a puppeteer must learn how to make use of strings and triggers to make their puppet dance, we in the digital age have to master a variety of controllers to effectively perform our tasks. We typically see controllers of mechanical devices with great functionality like a mouse with a scroll wheel, hotkeys, and side triggers. These function as great controllers in the context of 3D modeling. A more abstract type of controller is a MIDI controller which delivers data into a music program through a unique protocol designed for instruments.

But controllers don't always have to be hardware. To interact in a virtual space, it is not necessary to have today's most expensive VR controllers or a multi-component mounted tracking system in order to interact in a virtual space. Our bare hands can function as controllers using the right camera. Familiar objects like an old stuffed animal, or your favorite pen, can also be converted into interfaces that influence a virtual space. The re-purposing of familiar objects into virtual controllers is an enchanting augmentation, which offers the potential to use an object with great sentimental value to effectively perform complex tasks.

There are naturally problems with technology availability, cost, and processing power required to support unique objects for this purpose. But cameras are becoming a more cost effective tool for data representation and virtualization.

1.2. Measuring our world in RGB-D Data

Like most technologies, cameras are involving to support new applications and enable new uses for captured content. More and more cameras are being equipped with technology such as infrared lasers which are used to measure depth in an

1 https://direct.playstation.com/en-us/accessories/ps5

image. The 2D images that we see everywhere today are just an array of pixels each with a red green and blue color value (RGB). But by introducing depth data, these images become 3D - every RGB pixel value is paired to a corresponding depth value (D). An image in which every pixel, or point, has a color and the three dimensional location, is what we call a point cloud. Many people have never seen a point cloud but there are RGB-D representations of data all around us.

Autonomous driving is a field of innovation that heavily relies on cameras that collect depth data. They use a technology that primarily makes use of depth information - known as LiDAR - to capture the structure of nearby objects. The structures are then funneled into neural networks which classify the structures (more about this in Chapter 2). The networks make decisions based on metadata, such as stopping the car, steering left, adjusting speed, etc. LiDAR has also been used in the agricultural sector on autonomous drones that travel through acres of land and collect data about the color and size of crops to signal [2]. Figure 1.1 RGB-D data is also an invaluable tool in the construction and research industries as it enables a single camera to be sent to a location to record impressively high density scene captures, and the data can be imported and reconstructed in commercial softwares with relative ease [3].

Figure 1.1: Agricultural Drone using LiDAR for Precision Landscape Monitoring

2 https://www.blickfeld.com/blog/lidar-in-agriculture

3 https://bim360resources.autodesk.com/connect-construct/point-clouds-are-essential-to-the-construction-industry-here-s-how-to-maximize-their-value

In many contexts, the point clouds that are captured represent a single moment in time, and demand high functioning processors to retain information from one frame to the next. But recent technological advancements have led us to the possibility of RGB-D streaming, even for personal use. Therefore, I ponder, to what extent can point cloud data be used commercially to facilitate virtual interaction in our own familiar spaces? And how can our physical world influence a virtual world to be more familiar?

1.3. Hybrid-Space: What is Cross-Reality Environment?

The nature of content that we typically see in virtual reality is often not reflective of the environment around us. This is not a problem, as it is quite literally one of the primary objectives of virtual reality, to escape our immediate surroundings. However I strongly suggest that our personal environments are not ignored, but instead augmented through technology to be more preferable. Augmented reality is a strong tool that allows us to overlay virtual content into the physical world. But in the commercial context, this is typically done using an egocentric display that supports one main user.

If we can augment our environments using accessible commercial technologies, to feature virtual interaction, then our space becomes a "Hybrid-Space". This hybrid space allows for cross-reality interaction from the physical world to the virtual, and vice versa. Depending on the configuration of the environment, the physical and virtual spaces can be designed to operate together harmoniously through cross reality synchronization. This presents an abstract social perception of the environment, in which each side can react to and mirror its counterpart,

1.4. Proposal

As an engineer from an economically lower class, I found myself trapped in a social circle that frequently exposed me to highly customizable and advanced technologies that were never accessible within my budget. As a gamer, there are

many peripherals and custom controllers on the market that give a preferable experience to those from a stronger financial background. As a teacher, it is also discouraging to see many digital technologies that provide fascinating user experiences, but require high level computer science familiarity to get started, and even more to make full use of a product.

It is therefore the goal of this thesis to design a kind of tool, (whether a system, product, software, toolkit, or application) that makes use of relatively less expensive technologies, and allows users to engage in a new kind of interaction, regardless of prior technical knowledge. A primary objective is to normalize and popularize the use of RGB-D data for virtual interaction, and generate interest for new kinds of media and digital experiences. The secondary objective of this thesis is to propose a means through which users can augment their own environments into Hybrid-Spaces, increasing space functionality through the cross section of real and virtual worlds. Through UX design in a simple platform, higher level computer science concepts can be easily packaged, and accessed in just a couple clicks, or perhaps through the experience itself. In the case of this work, the unique interaction will consist of designing a new kind of virtual controller derived from live RGB-D Data, content displayed through a large scale visualization output, and an exploration of the quality of user interaction through this medium.

1.5. Thesis Structure

This thesis is divided into five chapters. Chapter 1 describes the background and user base that warrants a need for this thesis. Chapter 2 consists of an in-depth literature review, acknowledging relevant works and their limitations, methodologies, display types, and interaction scenarios which shape of design of this work. Chapter 3 outlines the design process and necessary functionality for this work to be effective. In Chapter 4 the prototype fabrication is documented, and evaluation metrics are discussed. Chapter 5 summarizes this thesis and poses some conclusions.

Contributions

The key contributions of this thesis are listed as follows:

- Acknowledged a lack of large-scale display point cloud interaction literature.
- Introduced a design concept for a toolkit to augment environments and enable cross-reality interaction.
- Created a graphical rendering technique that achieves localized data manipulation, circumventing the need for classification networks.
- Built a functioning large scale prototype.
- Highlighted significant features from multiple prototype iterations (although equipments used were not minimal-cost).
- Collected alpha stage feedback from system users.
- Established a validation measurement through survey design and observational data.

Chapter 2
Literature Review

2.1. Augmented Reality

Augmented Reality (AR) is the trendy technology that is used in our world commercially for applications ranging from education, to gaming, health care, and even remote industrial work [2]. In general AR can be described as a method of overlaying digital information on top of our real world perception, typically featuring computer vision, and a capable display [3]. It is not to be confused with Virtual Reality (VR) which has been around longer. The primary difference between the two is that AR functions as a lens to modify real world data, and VR typically hides the real world, bringing the user into a new cyber-space. However, the first AR systems did not achieve the effect that we come to expect today, and actually functioned much differently, in style and purpose.

Early Augmented Reality

Considered by some the first immersive and interactive AR system in 1992, *'Virtual Fixtures'* from Louis B. Rosenberg of the United States Air Force Armstrong Labs, was a system in which participants controlled an exoskeleton arm, using virtual perceptual aids, to accomplish a telemanipulation task [4]. This work demonstrated for the first time a benefit to human perception, and that operator performance could be improved by up to 70% through the assistance of an AR system (Fig 2.1). Following this work, the military, aviation and space industries boomed with works pertaining to computer vision assisted informatics and tracking systems, which normalized the technology until it began to shift to urban enhancement around the 2000s [5].

AR grew beyond its contribution to research and informatics, and became more

Figure 2.1: Virtual Fixtures Kinaesthetic Exoskeleton (1992)

commercially focused, with outdoor navigation and terrain visualization applications steering it into the public eye. It later evolved into a new and fun medium for entertainment, allowing users to view virtual content and music atop of the physical world [6]. Of course as computers began to shrink, so did AR systems.

The world of computational devices was becoming more portable - as a result, research turned to mobile devices to foster the next generation of AR and VR systems. Head Mounted Displays (HMDs) were popularized in media as the future of Mixed Reality (MR), alongside the rapidly-expanding internet, through which collaborative interactions would become a possibility [7]. MR became the umbrella term for works that use both AR and VR techniques for their functionality.

Modern Mixed Reality

Not only did MR bring much attention to display design, but also to virtualization of our physical world. Many systems are still being designed that work to take data, and objects from our world, and to replicate it in MR systems [8]. Furthermore, these methods of viewing have shown promising results such as enhancing the retention of information in education [9].

Today, a very popular subject in research is the use of interactive technologies to allow users to manipulate and control the content in their MR scenes. In the context of AR, this is achieved on the computer side by performing several core tasks as described by Freitas et al. [10]:

- Tracking the Environment

- Tracking the User
- Modeling 3D Objects

Following these tasks, a user can observe and technically manipulate virtual objects, however the quality of object interaction is still up for discussion. This list of objectives also fails to address a significant aspect of modern AR, which is real world object detection. There are many techniques that produce varying results in varying contexts. Moreover, the different methods of displaying AR content will strongly impact the perception of a system, which is what is explored in this thesis.

First we will observe the architecture of existing object detection systems, and evaluate their strengths and weaknesses in the context of AR design.

2.2. Detecting 3D Objects for AR Systems

2.2.1 Methods

To import content from our real world into an AR scene, we must first perform a process called object detection, which is the use of computers to track the position and orientation of any object, in a given space. Object detection is not only necessary to scan the items we wish to import into our scene, but in some cases it will also be used to scan the user's tools for digital interaction.

Depending on user preference, one may want to control their AR experience through multiple channels of input. For example, one user may want to interact using their bare hands, while another may prefer to use a controller. In other cases, a user may want to avoid interaction completely, and leave it to an autonomous system to perform AR object manipulation. In either case, there are many tools that achieve similar results, through different techniques.

For this thesis we will explore a few methods that have shown to be effective in recent literature.

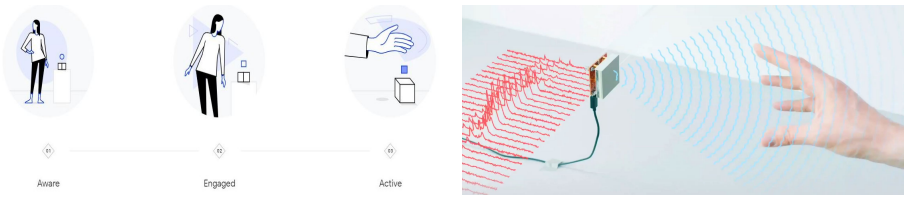


Figure 2.2: Google Soli - User Interaction Design

Electromagnetic Data

One intuitive way to track an object is to embed an electronic device into an object which sends data back to a central point. A common tool for this approach is the use of an accelerometer which can provide useful data such as orientation and location, but also requires a power source. A different approach that also uses electromagnetic data is a device like the Google Soli [1] which is a chip that detects human motion within a small range. Figure 2.2 shows the use case design of the chip as well as the hidden electromagnetic operation.

Other recent literature suggests that electromagnetic devices may be effective for tracking fine tuned human interaction in a small area but don't offer flexibility in tracking any unique object brought into the scene [11].

Infrared and Hand Tracking

Moving up the electromagnetic scale we approach the infrared (IR) spectrum which is commonly used in motion capture systems. By taking multiple IR cameras and an asymmetric pattern of IR reflective landmarks, we can calibrate and locate objects and even people, through a process called triangulation. This technique is great for rough estimation of objects in predefined spaces, but has a few drawbacks in the context of AR interaction. The first is cost, with these systems typically upwards of $10,000 [2]. Second is the precision of tracking, which isn't always great for fine tuned hand tracking in the frequent case of obstruction. Lastly is that the objects to be tracked must be predefined, and input into a system which

1 https://atap.google.com/soli/

2 https://www.vicon.com/hardware/cameras/

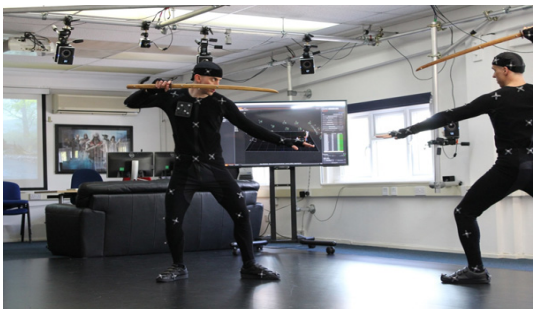


Figure 2.3: Vicon Motion Capture System

doesn't leave freedom to rapidly integrate new objects into a system. Figure 2.3 depicts two users in motion capture suits in an environment with a setup of at least 12 VICON motion tracking cameras.

Through computer vision techniques, we can improve on hand tracking using a product known as the leap motion controller [3]. This device has been used in recent literature in two key configurations: affixed to an HMD, or affixed to a stationary surface [12]. This method provides decent quality bare-hand-tracking for AR systems that use HMDs, and can be used establish areas where bare-hand-interaction provides significant data.

A weakness of this technology, though it performs it's role of tracking hands effectively, is that it it fails to offer simple configuration to track more abstract objects. One method that users have learned is to paste a picture of a hand onto an object, and the device can track it's location through deception.

Body Tracking via HMD

There are numerous products that piggy-back off the use of HMDs to perform user tracking. One such device is the VicoVR [4] which tracks the location of its companion VR HMD, and also performs full body tracking, using a peripheral sensor (Fig 2.4). The information from this system can be wirelessly transmitted to mobile Android and iOS devices via Bluetooth. This configuration presents an effective method to track users with full-body functionality, without the need for

3 https://www.ultraleap.com/product/leap-motion-controller/

4 https://vicovr.com/

Figure 2.4: VicoVR - Bluetooth Mobile Tracking System

excessive wiring and expensive cameras. An added benefit of body and gesture recognition, is that the inputs of human-activity can be used to synchronize, orient, and segment data through AR interaction techniques, as demonstrated in recent literature [13].

Touch Based Tracking

We have also seen modern works that perform object detection at a much higher resolution, using RGB-D data technology, and geometric contact sensors. In 2017, the Massachussetts Insitute of Technology demonstrated an efficiency in object detection by mounting two of these sensors onto a mechanical arm, and having it "feel up" small objects to collect data about their shape and color in virtual space [14]. Geometric sensors such as GelSight (used in the aforementioned work) have incredible detail, providing structure rigidity information as fine as 2 microns wide. They can serve as a great tool in the process of digitizing real world content, however the sensors are small and used on equally small objects.

Furthermore, this technique fundamentally operates using a semi-invasive principle of touching objects to recognize and determine their shapes. Lastly, though using this method for object tracking is accurate on small scale items, it isn't easily reconstructable nor portable, making it a hard technology to implement in an on-the-move user scenario. It would seem that using a geometric contact sensor would be more appropriate for a controlled space, but RGB-D data opens the door to another world of possibilities.

Figure 2.5: Microsoft Azure Kinect: RGB-D Camera

RGB-D Data

A colloquialism for RGB-D data, is 3D point cloud, which is a term commonly used in marketing to advertise color-depth camera capability. The most recent version of the Xbox Kinect, now known as the Microsoft Azure Kinect, features the same old body tracking API that the gaming community has come to know and love, but also fashions a new point cloud camera (Fig 2.5). Capturing depth and color, and streamlining it into computer software for analysis and reconstruction has spawned an entire subculture of modern literature.

Microsoft in their work *'KinectFusion'* perform dense surface mapping, via the new Kinect, by generating high resolution colored scenes in real-time using data streamed from the point cloud camera, and frame-to-frame tracking algorithms [15]. This work showed promise for seamless integration of real objects (and even spaces) into AR scenes, through reconstruction. Although, for more complex AR applications, there is a need for context of reconstructions, which in the computing world is known as metadata.

Metadata is important as it helps us classify objects, which has become a popular topic in the context of machine learning. There are many models that work to classify images of items such as chairs, tables, and other simple objects. But 3D point cloud classification through machine learning is relatively new field with much room for improvement in speed [16].

The context of an object is also significant to simulate it's behavior in an AR environment. Some objects deform, and these changes in shape among real objects determine the robustness of classification algorithms [17]. Simulating the interaction between deformed objects in virtual space also requires a very fine

tuned physics engine, and a very robust framework as demonstrated by Petit et al. [18].

2.2.2 Modern Algorithms

RGB-D data, using point cloud camera technology in the context of AR object detection and tracking seems highly practical and adaptable. The technology avoids spatial, cost, and weight limitations of the other methods previously discussed. Thus, point cloud research today has shifted into a direction of object recognition using unique algorithms for a plethora of scenarios and applications. Recent works such as *'ComplexerYOLO'* have put an emphasis on using convolutional neural networks (CNN) to address needs for AR applications like autonomous driving, using semantic (or segmented) point clouds for efficiency [19].

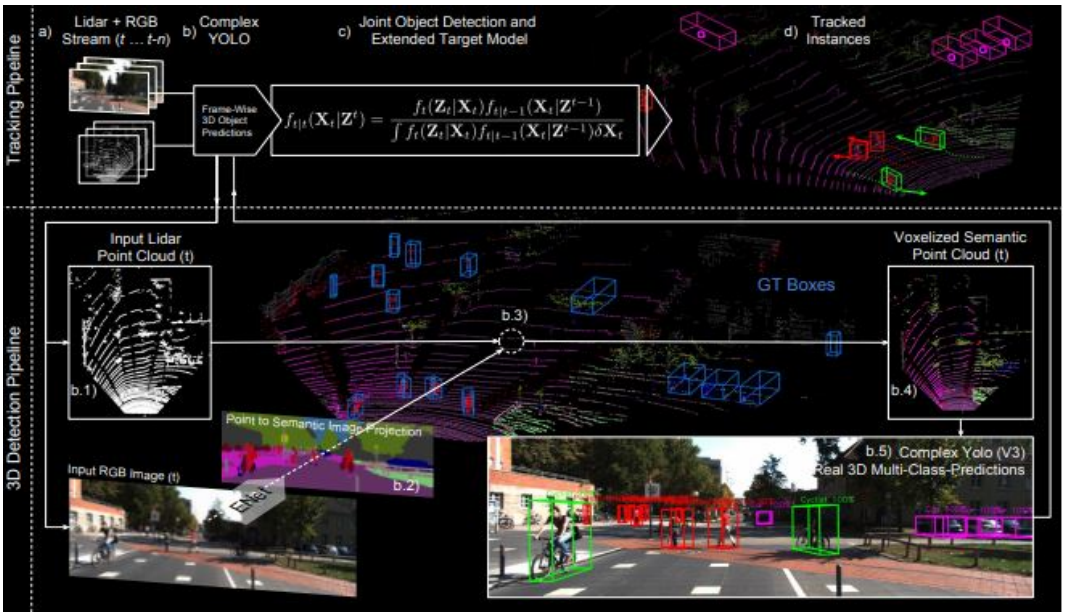


Figure 2.6: ComplexerYOLO Processing Pipeline

In another niche of the point-cloud-object-detection-subculture, researchers are working on designing frameworks that perform *object detection, feature extraction, data association,* and *more.* A competitively efficient framework from 2020, *'PointTrackNet'*, makes use of these modern algorithms and a simpler depth-information technology, LiDAR, in the context of autonomous driving [20]. An-

other framework, *'ZoomNet'* expands on regular 2D input, using pseudo LiDAR, to perform point cloud reconstruction, as well as pose estimation, and predictive algorithms. Though even the authors may admit that this 2D framework lacks in comparison to frameworks operating on 3D input, stating "...[it] can serve as a competitive and reliable backup for autonomous driving and robot navigation" [21].

AR has many more applications than just autonomous driving, in which the orientation of scanned objects becomes questionable. Typically in the context of driving, it's safe to assume that objects are normally perpendicular to the ground, but in interactive situations, this is not always the case. Hence, there is another niche of research in this subculture that has been producing algorithms that detect orientation from known object features.

Far prior to commercial point cloud technology, and even AR systems, object curvature and non-rigidity caused problems in computational vision that needed to be addressed using temporal solutions, or frame-to-frame comparison, and monocular to binocular image exploitation [22]. In 2010, a world where point clouds were growing in presence, Wang et al. proposed object curvature and point density as a technique to identify landmarks for object detection [23]. The advancement of landmark detection has led to orientation inference. In a recent work, Tencent in China [5] demonstrated object classification, alignment, and feature matching using sampling and landmark techniques as well as sperical voxels [24].

2.2.3 Computational Demands

Returning to the theme of AR object detection in the context of autonomous driving, a voxel based point cloud detection network (SVGA-Net) produced last year outperformed state-of-the-art RGB-D/LiDAR detection networks by 4-6%, but operates across four graphics cards which is a costly requisite for any computer software [25]. The network is also trained on a benchmark dataset for computer vision in driving, (KITTI) [6] which is well defined, but not designed to handle outlier inputs.

5 https://www.tencent.com/en-us

6 http://www.cvlibs.net/datasets/kitti/

Figure 2.7: Virtual KITTI 2 Training & Testing Dataset (Jan 2020)

Other object tracking works have achieved real-time object detection through point cloud data, tracking predefined objects through point-matching and interpolation between frames [26]. Frame based approaches are commonly used in applications where point clouds are being rendered in real-time, but again lack the flexibility to handle outlier objects that deform. Training a network to learn about a new object in a running instance isn't an easy task if it isn't sure what defines the object.

Recent point-cloud learning algorithms have been broken into these two main categories, *voxel based learning* and *point based learning* and come with their own weaknesses [27]. Voxel based algorithms account for a lot of space that is often unused, and require more computational power, while point based algorithms use up much more memory which isn't always feasable on mobile devices for the context of AR. One network, Grid-CGN [28] limited the amount of input points to just over 80,000 saving significant time from data-structuring to achieve a inference speed of up to 50 fps (identification averages of between 15-40ms as opposed to competitors averaging 80-200ms as shown in Fig 2.8).

Rendering point cloud data, tracking objects, and creating meshes from data, all to be viewed in live virtual space is generally a frame rate killer. The goal in the context of interactive AR would be to achieve a fast-acting classification network that can learn new deformable objects, and provide tracking along with pose estimation for digital simulations. Implementing these frameworks, on the other hand, in an entertaining fashion for mobile applications also takes great skill, and foresight to design an experience that isn't just a copy of another popular app. This is not only key for AR, but recognition and pose estimation of free-form objects is also significant for autonomous robotic manipulation [29].

Input (xyz as default)		ModelNet40 OA	ModelNet40 mAcc	ModelNet10 OA	ModelNet10 mAcc	latency (ms)
OA ⩽ 91.5						
PointNet[29]	16×1024	89.2	86.2	-	-	**15.0**
SCNet[44]	16×1024	90.0	87.6	-		-
SpiderCNN[47]	8 × 1024	90.5	-	-	-	85.0
O-CNN[40]	octree	90.6	-	-	-	90.0
SO-net[22]	8 × 2048	90.8	87.3	**94.1**	**93.9**	-
Grid-GCN[1]	16×1024	**91.5**	**88.6**	93.4	92.1	15.9
OA ⩽ 92.0						
3DmFVNet[3]	16×1024	91.6	-	95.2	-	39.0
PAT[48]	8 × 1024	91.7	-		-	88.6
Kd-net[15]	kd-tree	91.8	88.5	94.0	93.5	-
PointNet++[30]	16×1024	91.9	**90.7**		-	26.8
Grid-GCN[2]	16×1024	**92.0**	89.7	**95.8**	**95.3**	**21.8**
OA > 92.0						
DGCNN[41]	16×1024	92.2	90.2		-	89.7
PCNN[2]	16×1024	92.3	-	94.9	-	226.0
Point2Seq[24]	16×1024	92.6	-			-
A-CNN[16]	16×1024	92.6	90.3	95.5	95.3	68.0
KPConv[36]	16×6500	92.7	-	-	-	125.0
Grid-GCN[3]	16×1024	92.7	90.6	96.5	95.7	**26.2**
Grid-GCN^{full}	16×1024	**93.1**	**91.3**	**97.5**	**97.4**	42.2

Input (xyz as default)		OA	latency (ms)
OA < 84.0			
PointNet[29]	8 × 4096	73.9	20.3
OctNet[31]	volume	76.6	-
PointNet++[30]	8 × 4096	83.7	72.3
Grid-GCN$_{(0.5\times K)}$	4 × 8192	**83.9**	**16.6**
OA ⩾ 84.0			
SpecGCN[37]	-	84.8	-
PointCNN[23]	12×2048	85.1	250.0
Shellnet[50]	-	85.2	-
Grid-GCN$_{(1\times K)}$	4 × 8192	**85.4**	**20.8**
A-CNN[16]	1 × 8192	**85.4**	92.0
Grid-GCN$_{(1\times K)}$	1 × 8192	**85.4**	**7.48**

Figure 2.8: GridCGN Accuracy and Speed Comparison Tables

2.3. Viewing 3D Objects in AR Systems

2.3.1 Small Scale Displays

Wearable Displays

Although there is a serious drawback to using mobile devices for computationally demanding applications such as AR, it has not significantly impacted the popularity and demand for mobile HMDs. There is a large market of manufacturers who design AR HMDs with varying capabilities from inward-outward tracking [7], to transparent displays [8], to simplistic smartphone housing [9] (Figure 2.9).

This is understandably so, for multiple reasons. Inexpensive wearable AR displays have been shown to give a sense of presence to users in remote spaces which allows them to share the same experience [30]. Even without network connectivity, near-eye AR displays "[enrich] real-world visual experiences with digital content ... enhancing the human experience and task performance..." as Dunn et al. puts it [31]. Although not entirely critical to a full AR experience, HMDs

7 https://www.oculus.com/quest-2/

8 https://www.microsoft.com/en-us/hololens

9 https://mergeedu.com/headset

a) Oculus Quest 2 b) Microsoft Hololens 2 c) Merge AR/VR Headset

Figure 2.9: Commercial High-End AR Headsets

can be simply described as *small tools* that let users visualize *large things*.

Other Small Scale Displays

A point of high interest is the focus on human cognition while operating using AR. It is important to address other relevant works and use cases of small-scale devices that operate through vision manipulation to *perceive* AR constructs. One such alternative to HMDs are an emerging technology known as retinal projection displays (RPDs). Peillard et al. found that RPDs provided more accurate depth estimation of virtual objects by users, than what was demonstrated using typical Optical See-Through displays (OSTs) [32].

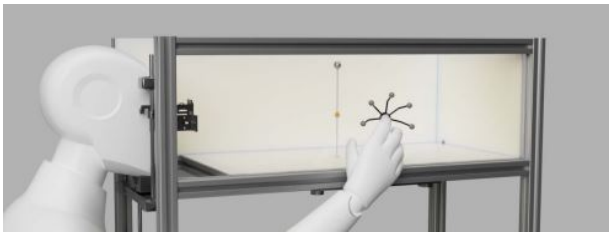


Figure 2.10: Retinal Projection Display - Peillard et al. (2020)

Mobile applications that perform photogrammetry and make use of LiDAR can make realistic, dense, 3D reconstructions of physical spaces. A challenge in achieving interaction through these applications, is the time it takes to process and render data [33]. It is another challenge entirely to insert new virtual objects into a captured scene to perform augmentation. Zhang et al. designed a robust system

for inserting virtual objects into monocular videos [34], but even their state-of-the-art work has severe limitations in highly dynamic scenes, or even scenes in which the camera does not remain perfectly vertical.

2.3.2 Large Scale Displays

Large scale displays, on the other hand, aren't as severely restricted by computer size, and may have entirely different impacts on users. Immersive environments are commonly portrayed with large screens that surround a user, like theatres, gaming monitors, and curved displays. By using light manipulation, and optical properties, there is another genre of displays, inspired by projectors, Light Field Displays.

TV Panels

Light Field displays make use of a variety of different techniques for different purposes. Classical approaches using screens that feature a display pixel matrix, suffer limitations such as limited viewing angle, insufficient depth, and low resolution. The recent work of Nam et al. attempts to address the resolution issue using blur effects in an autostereoscopic 3D display as opposed to the more popular technique of using light polarization [35]. Their results demonstrated better resolution using a very experimental and emerging genre of screen technology.

Volumetric 3D Displays

A different approach to achieve large scale visualization is through volumetric displays [10]. These kinds of displays are aesthetic in their holographic nature, and can be perceived from many angles. However, they're quite bulky and the ratio of display area to hardware makes the utilization of these displays in large, shape changing environments, rather impractical.

10 https://voxon.co/

Figure 2.11: Voxon Photonics Volumetric Display

Holography - The Vision

A highly cited paper from 2006, *A Large-Scale Interactive Holographic Display* [36], features one of the most promising applications of holographic screens. In this work - using projection technology - multiple freely moving naked eye users are able to observe content on a extremely high resolution (50 million pixel) display. This brings us to a realm of literature that focuses on projection based light-field displays.

Projection Mapping

Projection mapping is an artistic method that has become widely popularized in pop-culture and provides a high degree of immersion through the prinicple of transforming surfaces through light manipulation. Figure 2.12 displays the project: Lightwing 2 [11], which features stereoscopic projections and provides some tactile data through the handles on the display, which gives the user the experience of "piloting a space flight". Figure 2.13 shares another popular work: "Le Petit Chef" which is an interactive dining experience in Belgium that uses downward projection mapping onto tableware to tell stories and convert surfaces into semi-interactive canvases [12].

11 https://ars.electronica.art/outofthebox/en/lightwing2/

12 https://skullmapping.com/project/le-petit-chef/

Figure 2.12: Lightwing 2 - Interactive Installation

Figure 2.13: "Le Petit Chef" - Projection Mapped Dining Experience in Belgium

Figure 2.14: Bangkok Projection Mapping Competition 2021

Figure 2.15: TeamLab Borderless Exhibition in Japan

Projection mapping in public venues such as the annual Bangkok Projection Mapping Competition in Thailand [13] and the TeamLab Borderless exhibition in Japan [14] are aesthetically appealing to a high degreee, however a common trait in recent projection mapping works, is the lack of user input. Not always large scale, projection mapping assumes smaller configurations such as the illuminated drum[15] featured in Figure 2.16 which can give new meaning through visual augmentation to unique objects such as musical instruments.

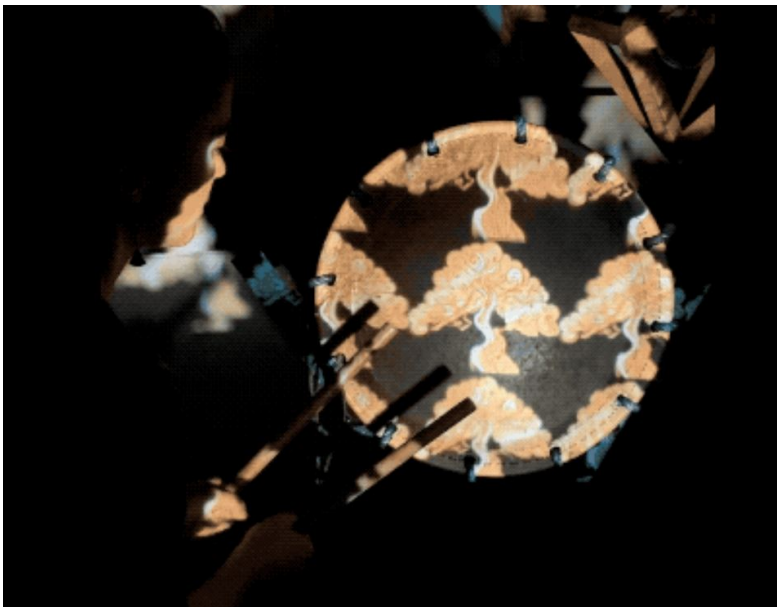


Figure 2.16: Taiko x Monk Beatbox x Mapping - Remy Busson

Projection Displays

Light Field Projection Displays have been around for many years, and have been used digital media for many-person viewing, as well as single-person viewing. The common design is for a projector to illuminate a type of screen such as the

13 https://www.bangkokdesignweek.com/en/program/87121

14 https://borderless.teamlab.art/

15 https://remybusson.com/projection-mapping

following prototype by projection media specialist: Scott Allen [16] (Figure 2.17).

Figure 2.17: "Polid Screen Test 002" - Scott Allen

In recent years a goal of many works has been to use light field displays to generate a true three-dimensional viewing experience. In the work of Ni et al. [37], a 360° display system was achieved using 360 projectors onto a cylindrical screen, which perhaps may have been a bit overkill on visual quality. As discussed previously, there are many segments of the data rendering pipeline that can be improved to achieve higher quality resolution, such as preserving streaming bandwidth [38]. If a display achieves the maximum resolution quality for any number of users, the question in the AR context then becomes: *"How do we make it more interactive?"*

In RePro3D, a system was designed that produced haptic feedback when interacting with a holographic-projected object [39]. A number of experience augmenting works since then have been produced that operate through wearable devices. It can certainly be contested that supplemental technology can increase our responsiveness and impression of a virtual experience. However, Yu et al. argues that "information received by human eyes accounts for more than 80% of all external information received by humans" [40]. They further suggest that modern light field reconstruction systems can realize scenarios in which "the viewer does not need to wear additional equipment ... and the experience is comfortable, without fatigue". This is in regards to visualization technology such as HMDs,

16 https://scottallen.ws/log/blog/amid-screen-polid-screen-test_001/

but wearable devices that restrict body movement, or are significantly weighted, do indeed cause fatigue.

So, how do users interact with large scale displays in a way that isn't exhausting?

2.4. Interactive Environments using AR

2.4.1 Virtual Controllers

The cornerstone of comfortability in AR, is the mode through with the user interacts with the perceived data. Whether a user can sit by and watch as new information streams in their sight, or if they are encouraged to reach out and "touch" their egocentric digital content, the devices that control, initiate, and facilitate the experience through digital processing are the Virtual Controllers. They can take the shape of conventional gaming controllers, or they can appear more abstract in forms such as a vision system, haptic devices, embedded systems, and even our own bodies.

For starters, through computer vision applications, we can transcend from the barbaric method of attaching wired wearables to our bodies and rely on external tracking systems to monitor our movements. Cho et al. [41] propose a system that uses a generative adversarial network to track human hands whilst interacting with physical objects. The design allows bare hand interactions with tangible objects that circumvents occlusion problems in AR contexts as shown below in Figure 2.18.

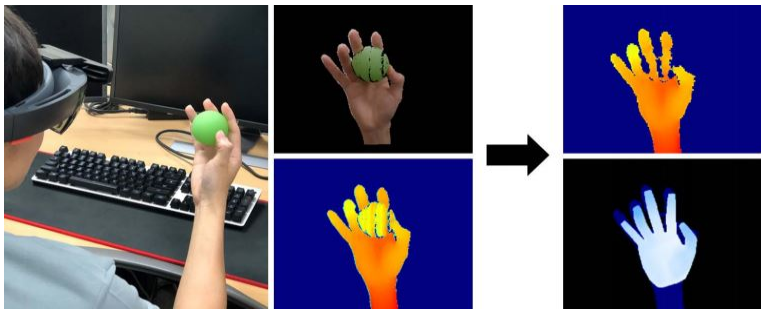


Figure 2.18: Bare-hand Depth Inpainting - Cho et al. (2020)

Another interesting application of RGB-D data, is the *RGB-D-Event* camera

system designed by Dubeau et al. which incorporates additional temporal data into a virtual system to detect motion based events [42]. In this work, an arbitrary object can be converted into a controller under the assumption that it can move. This kind of design can shift the role of movement from a user onto a different entity entirely, if exhaustion is too much of a concern.

Virtual Agency

The use of external systems to control, manipulate, and exist in the digital space has spawned a new field in which an important checkpoint is to achieve a sense of agency, or vaguely put: "the feeling that you are really in control through your controller". By shifting the tools of control from conventional controllers to our own bodies, AR users can experience more fulfilling scenarios, and reach this agency. Modern techniques such as inverse-kinematic avatar movement, "[increase] the sense of embodiment and the sense of spatial presence" in virtual space [43]. But the *limitations of egocentric design* strike yet again - as a lack of complete spatial data representation can lead to incorrect depth assumption and thus obscure virtual presence. Hence, it has been a challenge to design new methods of environmental augmentation to fill in the gaps where HMDs lack sufficient data, like in the *Augmented Mirrors* project [44].

2.4.2 Interaction with Immersive Displays

The Role of Media

Stemming from the many works that seek to create new interaction scenarios, one might paint the future as a landscape in which interactive displays are as commonplace as the smartphone. To a certain degree, that prophecy is already fulfilled as many smartphones today function as interactive displays. Although not used for outdated user scenarios like barcode scanning as depicted by Billinghurst [45], wearable devices are becoming more commonplace, like the smartwatch. Less invasive and much more lightweight than other larger wearables, the smartwatch gained traction through stylish fashion, and media popularization.

By this principle, if large scale interactive media becomes as obtainable and dependable as personal computers, then our culture may shift to develop more

digitally enabled spaces. What do immersive displays have to offer then?

Holography Then & Now

In an early work by Balogh et al. [46] a holographic environment was presented in which multiple users could interact continuously with a 3D scene, from their own point of view; which is called observer independent parallax. As exciting as it was to finally see holography (as idealized in pop-culture like Star Trek) in a manipulatable scenario, the work had many limitations - primarily space, but also viewing angle limitations and early generation gesture control.

Modern works have dusted these limitations, such as the interactive holographic display demonstrated by Takenaka et al. [47] from just this year. Using lasers, light refraction, and motion sensors, this work allows users to draw and even erase holograms using their fingertips, in real time (Figure 2.19). This type of creative and depth-driven interaction is highly empowering, especially for users who don't want to be outfitted with wearable devices. The environment is so controlled and precision focused, that it's difficult to scale it for larger applications. Let's observe some different interaction designs that empower users in a similar way, without the restrictions of space.

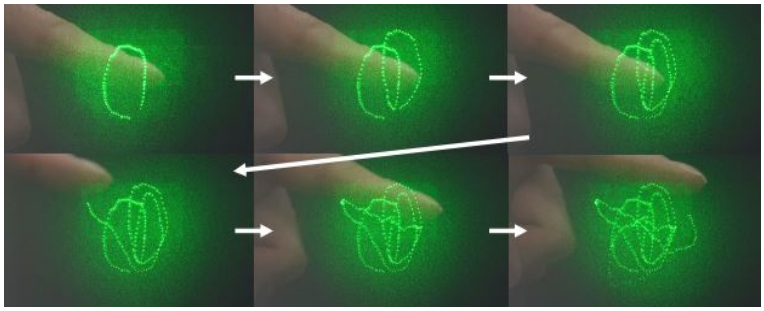


Figure 2.19: Interactive Holographic Display - Takenaka et al. (2021)

2.4.3 Variable Scale Interaction Designs

It would be fantastic if there was a universal AR system in which users don't have to worry about boundaries or spatial limitations. Then, theoretically, a user could walk between different contexts freely and start interacting with anything

they wish through an egocentric design. A hybrid-space approach could easily distinguish areas in which AR interaction is enabled but these spaces would need to be built and configured beforehand which would take time and money. Finally we will look at some egocentric interaction designs from a few recent works.

Collecting Physical Data

Project Zanzibar [48] is a portable mat that can uniquely identify tangible objects, communicate with them, and sense a user's touch as well as gestures above and nearby the mat (Fig 2.20). This is an interesting platform as it is not a wearable device, but it is something portable and smart that can be brought into any environment to facilitate AR interaction. While not specifically a visual application, the project does provide a user-centered experience without the need for a wearable HMD.

Figure 2.20: Project Zanzibar - Villar et al. (2018)

Processing Virtual Data

Alternatively using the HMD approach, there could potentially exist a "perfect, universal head mounted, computer vision AR system" that could freely enter and exit any environment and create meaningful interactive experiences. Aside from object recognition this task is already exceptionally difficult due to the common problem of occlusion in AR displays. GrabAR [49] combats this issue by predicting virtual object location and collision using a convolutional neural network. For the sake of universal compatibility, this shows that a device could be pre-equipped with an arsenal of supported objects that can be used to simulate realistic and

digitally responsive controllers. Figure 2.21 depicts the native overlay of an AR system (a), as well as the data retrieved from a depth sensor (b), and skillfully merges the two inputs into a final properly flushed result (c).

Figure 2.21: GrabAR - Tang et al. (2020)

As we have seen designs for collecting physical data, as well as virtual data from a user, how can those two crossroads intersect in a useful and entertaining way? One method is through Virtual Effects such as the publicly available AKVFX Unity Plugin shared by Keijiro Takahashi demonstrated in Figure 2.22 [17].

Lindlbauer et al. [50] presented Remixed Reality which was a project that used the Microsoft Azure Kinect to capture information of a live scene in the form of point cloud data. They then used a simple VR headset and controller to visualize, segment, duplicate, erase, and even translate data. The work then even made it possible to pause time so that the user could change their viewpoint at any desirable moment (Figure 2.23). This design that embodies the potential of

17 https://github.com/keijiro/Akvfx

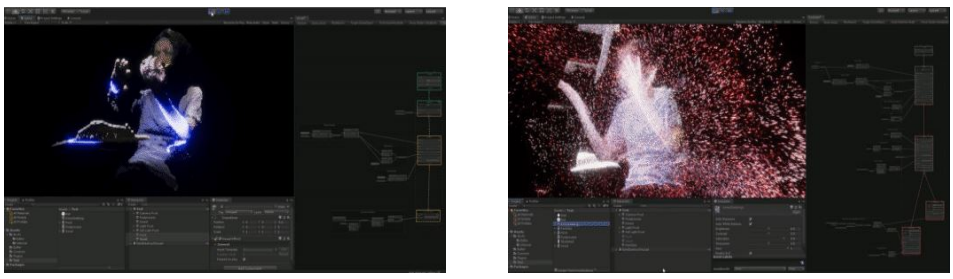


Figure 2.22: Akvfx Unity Plugin - Keijiro Takahashi

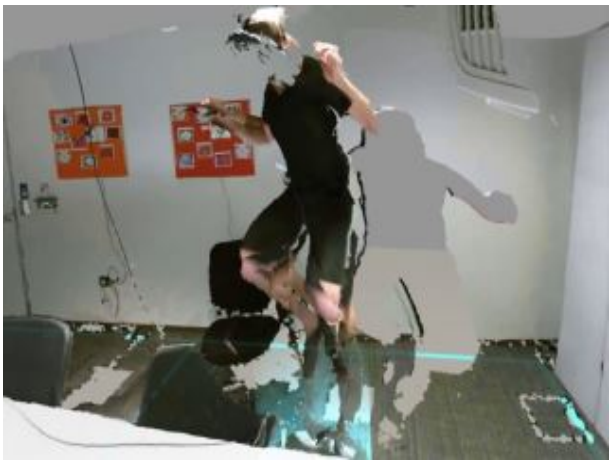


Figure 2.23: Remixed Reality - Microsoft (2018)

physical and virtual data manipulation is quite strongly related to the concept design that I will propose in Chapter 3.

2.5. Summary

In this chapter we have briefly seen a history of augmented reality works, and how they have impacted research directions over the years. Starting with immersive headsets aimed at facilitating telemanipulation, to informational display systems, AR evolved to become highly commercially obtainable with applications ranging from entertainment, to data segmentation, and autonomous driving. We have seen that in entertainment, there is a strong correlation between enjoyability, immersion, and interaction.

There are many modes of input for digital systems, and one that shows much promise is the computer vision approach of using point cloud data. There is a busy scene of research groups seeking to optimize object detection and classification in real time using point cloud data, but the methods require industrial level equipment, and complex neural networks. These limitations make it difficult for average consumers to see, interact with, and make use of point cloud data.

We also viewed different methods of observing objects in augmented reality, ranging from a small-scale wearable displays to large scale light field displays. Lightfield displays offer multi-user parallax, strong resolution quality, and aes-

thetically pleasing design - as influenced by media. Interaction with holographic systems is quite limited in modern works, and there are not many among these works that incorporate live streamed point cloud data into the design.

Other modes of input into hybrid-spaces were considered for the proposed system design in Chapter 3, but ultimately resembled peripherals that didn't contribute to the main RGB-D based interaction. Finally a work that made use of the Microsoft Azure Kinect RGB-D camera, and demonstrated its functionality in a game engine through the output of an HMD. Hence, this thesis differs from the previous work in terms of output type, and data interaction method.

Chapter 3
Concept Design

3.1. Concept

The current state in the history of AR is exceedingly focused on mobile technologies and delivering experiences that are presented to a single viewer. As a designer and researcher, I would like to challenge the norm and propose a redirection of current technology to facilitate AR interaction in familiar spaces that don't operate with the hardware limitations of mobile computers. This idea is largely inspired by the popularization of household gaming consoles, which give players a variety of digital experiences in the comfort of their own room. Projects like Illumiroom [51] have further improved television gaming and movie experiences using projection technology to expand perception using projector based environmental augmentation. A significant factor in the perceived quality of recreational digital experiences is the quality of the display. A smaller screen leads to difficulty in observing content and a lack of received information. However large displays are not an option for everyone due to the cost barrier and perhaps shape in some cases.

There are of course many other interesting ways that we can perceive digital content that aren't necessarily visual in nature. For example haptic feedback and spatial audio are great mediums to enhance a user's experience, but aren't always comprehensible without an accommodating visual representation. We can achieve a better sense of reception to a virtual world if our perception of that world is understandable and particularly entertaining. But, in the opposite direction, AR technology also provides an opportunity for us in the real world to influence the virtual world. Previous works in Chapter 2 showed us that RGB-D data can be used to import real objects and places into virtual space, but processing

requirements for classification networks made systems complex and slow. Works featuring large scale holographic displays exemplified the immersive potential and extreme visual resolution offered by projection light field displays. In interactive designs, data is typically taken from a physical space, processed in a virtual space, and then displayed to users in an egocentric fashion; without a continuing path for virtual output to influence the physical space.

RGB-D Data in Cross-Reality Environments

What if we turned this path into a loop? What if we could design a space in which the virtual world influenced the physical world which would then re-influence the virtual world and so forth in a closed feedback loop? Then the facilitating space would become a cross-reality environment. In this case neither side (physical or virtual) is more justified than the other, but they act more like mirrors; each contributing to the overall experience. The objective of the spatial design is for the physical space to be intimately connected with a virtual counterpart, not only mirroring each other, but actively influencing each other. To users who inhabit a space where both the physical and virtual is easily accessible, the familiar places like a bedroom could be "expanded" in a sense, by generating their own virtual counterpart. Qualities of the virtual space can be shown in the physical space, and elements within the physical could be imported into the virtual space. This type of system augments the perceived volumetric capacity of a physical space, and can encourage users to be more familiar with the virtual counterpart; accessing and traversing the virtual space frequently.

3.2. Research Goals and Direction

3.2.1 Research Goal

The research goal of this work can be broken down into two main points. These two points set the direction for the design of this work and outline the societal contribution of this thesis:

- To address the lack of RGB-D data driven virtual interactions in media by demonstrating new and interesting interaction techniques made possible

using a single point cloud camera as an input.

- To encourage the normalization of relatively low-cost spatial augmentation using AR techniques through a software toolkit that doesn't require heavy computer knowledge to operate, and DIY holographic displays composed of PVC pipes, projectors, and plastic screens.

3.2.2 Increasing Cross-reality Responsiveness

This thesis will also explore the sociological impact of a new paradigm of user presence - virtual space presence, synchronized with the presence from within the physical space that the user operates in. Since hybrid spaces inherently in design should be accommodating to many different tracking and visualization scenarios, my goal is to measure the responsiveness of the proposed tool, and optimize it to achieve a holistic and enjoyable, high speed interaction scenario. Perhaps a user may want to interact with themself or a friend more than the content of a virtual space, and thusly the space should accomodate multiple users even if they are in remote locations.

I propose to shift the computational workload onto the environment, and use RGB-D data as input for interaction for bare-body experiences. Therefore, I will design an environment that seeks a single user inside of it, tracks their movements, interprets the data in a virtual space, and reflects into the physical space a continuous, real time series of effects that foster imaginative freedom. My intent is that the application of this system will spark interest of designers to create new RGB-D data-driven interactive techniques to be used in the contexts of gaming, public exhibitions, data management, and telexistence.

3.2.3 Proposed Interaction Design

The interaction design of this work is heavily inspired by projection mapping works that transform large areas and surfaces into visual artworks. The design of the system will first encompass capturing the color and depth data of the user in the physical space, and replicate that data in the virtual space counterpart.

The virtual space shall act like a mirror of the physical, however, since the virtual space is not bounded by the limitations that we encounter in the real world,

it is free to ignore properties such as gravity and object rigidity. Content in the virtual space can be duplicated, reshaped, recolored, shattered, and manipulated through intangible gesture based interactions. A typical design of localized projection mapping configurations is a unique object that becomes repurposed into into semi-interactive canvas like the *illuminated drum* aforementioned in Chapter 2, but the source of digital influence remains unclear in this demonstration. Is the drumstick causing changes, or the tension in the drum? Is it a microphone in the environment? Thus the user in the proposed hybrid-space should have a clearly defined "virtual controller" that will serve as their primary mode of interaction control.

It is key to this thesis that the interaction does not stop in the virtual space. To facilitate a two-way interaction, there must be a conduit that represents the virtual content in the physical space. The conduit in this work will be a holographic visual display, as it is believed that visual content can be perceived at a higher rate than through other senses, such as auditory or sensational feedback. Furthermore, the holographic display will contribute to a sense of augmented reality, and may present some interesting depth-based interaction scenarios.

Finally, the design assumes that digital content perceived in the physical space will encourage the user to make micro-adjustments to their movements or posture with the expectation to observe new preferable and varying content. This is referred to as the "closed feedback loop".

3.3. System Architecture

3.3.1 Input: Data Capture using Kinect

From a technological standpoint there are limitations to the quality of input that the proposed system will operate with. This is in part due to budget constraints, but also due to the capability of currently available commercial products (which have recently been in higher demand and lower stock in the COVID-19 situation). However, this coincidentally is in line with the theme of low-cost commercially available products for DIY implementation.

For input, the system will use the Microsoft Azure Kinect [1] which features an RGB-D capable camera as well as an API that supports body tracking data. The Kinect was chosen for its dual modalities in both point cloud streaming and body tracking, as well as it's compatibility with the Unity Game Engine [2] as demonstrated in relevant works [3]. The streamed data also has an output rate of 30 FPS, which technically makes it a viable tool in cinematic applications.

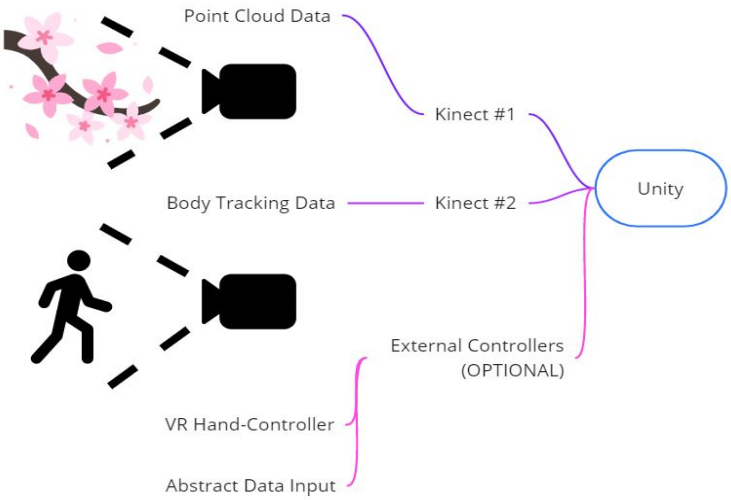


Figure 3.1: Input Data Pipeline

Since the Kinect will perform as a tracking system in the hybrid-space, it is customary that it will be affixed to the edge of the space, typically in a high corner where it will capture the most amount of data. The system will also explore the impact of virtualized objects, which is why a second Kinect will also be used inside and outside of the space where it's purpose will be to capture data of a predetermined object. A general simplification of these two inputs, as well as the compatibility for others using this toolkit, is demonstrated in Figure 3.1.

1 https://azure.microsoft.com/ja-jp/services/kinect-dk/

2 https://unity.com/

3 https://rfilkov.com/2019/07/24/azure-kinect-examples-for-unity/

3.3.2 Processing: Cross-reality synchronization

The virtual space is essentially a sandbox that will be used to design different kinds of contents that should be aesthetically appealing in the physical space. It is quite literally a space where the possibilities are limited only by our imagination and - technically speaking - our programming ability. Since the output of the system can be abstract and encompassing of many different kinds of visuals and effects, it is at this point where this concept can be categorized as a virtual toolkit.

This toolkit will process point cloud data in a game engine (in this case Unity) and should perform a variety of data manipulation techniques. One particular manipulation that I found very interesting was the concept of data multiplication. This means that a single point cloud can be piped into the toolkit, and in the virtual space we should be able to see the same data in multiple places from multiple perspectives. This presents the user a benefit of multi-vision that we generally lack in the physical world.

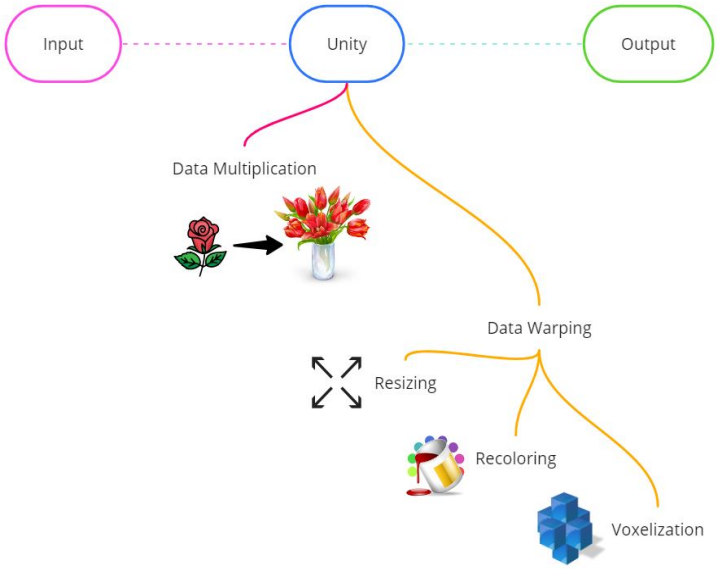


Figure 3.2: Data Processing Pipeline

Recent advancements in Unity's graphical rendering pipeline have also made it possible to apply virtual effects to our data. These effects range from recoloring, to reshaping and resizing, to voxalizing, and even to warping the RGB-D input

(which can consist of millions of input points). Figure 3.2 visualizes the different routes of virtual augmentation that the input data can take through this toolkit. This work will perform visual augmentation across millions of particles in the scene using VFX templates that are packaged similarly to the AKVFX Unity Plugin shared by Keijiro Takahashi mentioned in Chapter 2.

In terms of choreographing interaction, the toolkit will be designed to receive two modes of data which will facilitate physical to virtual interactions as well as virtual to virtual interactions. To replicate physical objects point cloud data should be sufficient input. For purely virtual interaction, any preferable virtual controller works, as well as the Kinect's body tracking API to track a user's presence in the physical space.

3.3.3 Output: Multidirectional Holographic Display

What will be viewed in the physical space? Generally speaking most digital experiences are contrived from a user and one main display. The display of course may be extended, curved, reshaped, or mounted in a clear and inconspicuous location, to achieve a better sense of immersion. However in the design of this work, there will be no single, *main* display.

The toolkit will be used in a scenario where there are multiple displays that will serve as windows into the digital space. This means that as a user freely meanders the physical space, they will also simultaneously be navigating the digital space. The displays can be installed anywhere in the space, in any direction, and should emphasize the digital expansion of the environment.

What kind of displays will be used? Are they just typical LED monitors? For this design, large LED monitors are not only exceedingly expensive, but are also heavy and difficult to install in DIY situations. There is also not much novelty in using displays that lack the element of depth. Therefore this toolkit will be used in a setting with multi-directional holographic displays, which will facilitate an AR experience, and can be set up with relative ease and inexpensive materials - as demonstrated in the work of Scott Allen, referenced in Chapter 2. A simplified diagram of the output pipeline can be seen in Figure 3.3.

It will also be interesting to explore the impact that large-scale displays will have on users. My hypothesis is that if users are able to move freely and unencumbered

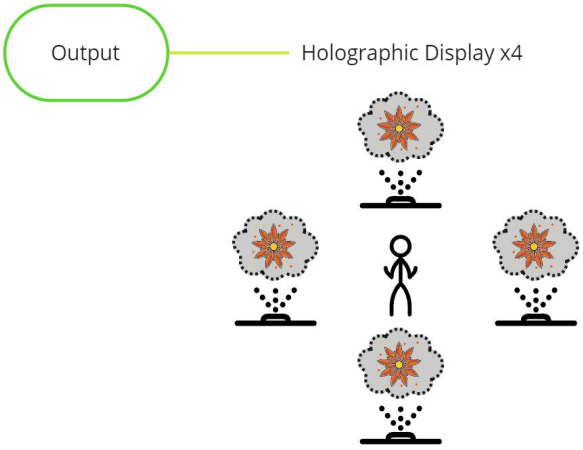


Figure 3.3: Data Output Diagram

in a large space to achieve visual stimulation from human-sized content, then this will promote high user activity in the system, which would contribute to an overall sense of immersion.

3.3.4 Interactions: Localization of Virtual Interactions

Likely, one of the most imperative points in this thesis is the design of interaction, in which I would like to propose something unexplored, novel, and valuable that can serve in a variety of applications. There are two key paradigms that will be exhibited through this toolkit which feature the interaction between a user and an object, and an object with another object. We will define interaction here very strictly to "a collision between two beings that results in a realistic, simulatable, and intent-driven outcome."

Using this self-defined criteria for an interaction, I have composed a flowchart below that models some bi-directional, and single directional interactions, with a highlight on the interactions that are possible through AR techniques inherent in this work (Figure 3.4). This graph is used to signify the expected events that will occur using the toolkit - and per event: which being in the interaction should exhibit a response. The proposed relational graph encompasses many typical activities between users and objects in the contexts of gaming, telemanipulation, navigation, and physics simulation.

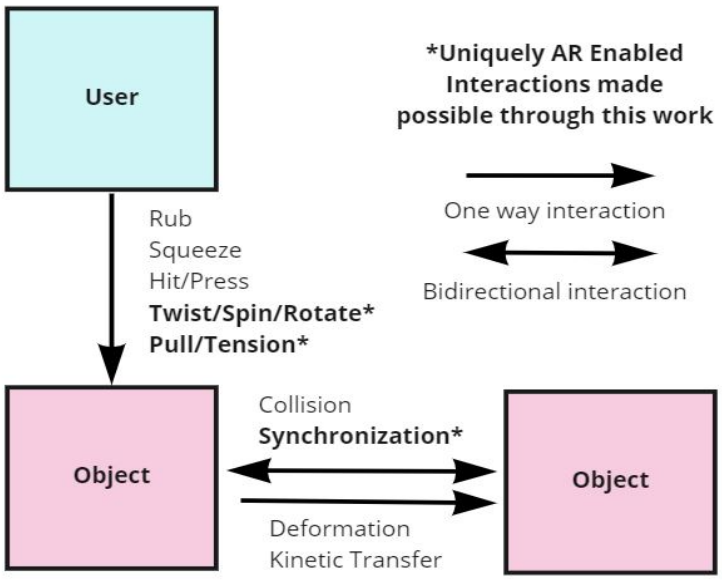


Figure 3.4: Human-Object Interaction Flowchart

Although, achieving this output with point cloud data as an input is much easier said than done. To even determine what beings exist in the scene, will require not only body tracking for users, but also a robust and fast functioning object classification method. It is for this reason that many works featuring digital effects in point clouds typically render the same effect across the entire scene of point cloud data.

In fact in these types of works, where there is no specified trigger event, visual effects are usually played on loop. Although this can be aesthetically pleasing, the system over time loses its originality and users have no way of influencing their environment to generate new, personalized experiences. In this thesis I want to avoid this norm, as it feels that the user can contribute no localized input. This toolkit will feature a *homebrew method* that should isolate subsections of the scene and apply the desired interaction effects using a depth based, identification-and-tracking pipeline built from the ground-up.

Chapter 4
Proof of Concept

4.1. Overview

In this chapter I will describe in detail the toolkit prototype that I built using the design specifications as outlined in Chapter 3. I will describe on a technical level the assembly process, my unique contribution to the data processing pipeline, and early stage user feedback. The name of this prototype is *Bridged Reality*.

4.2. Technical Implementation

4.2.1 Kinect, Holographic Displays and Unity

First and foremost it should be made very clear that this prototype is designed to augment a familiar space. This implies that there should be a sufficient area of space ready for use before building. Since some technologies such as the HTC Vive require a minimum 1.5x1.5 m^2 area to operate in, I began designing a space with at least double these dimensions for compatibility purposes. On paper the physical space would cover 4x4 m^2 of flooring to accommodate multiple people and potentially any large objects that might be inserted later (Figure 4.1).

Physical Setup

With these dimensions assigned, it was time to decide the placement and orientation of the holographic displays. For the most simple and geometrically convenient setup, I opted for a square space with four displays at each edge facing the center. I followed a guide (written by the Holo-Screen vendor) that listed out parts

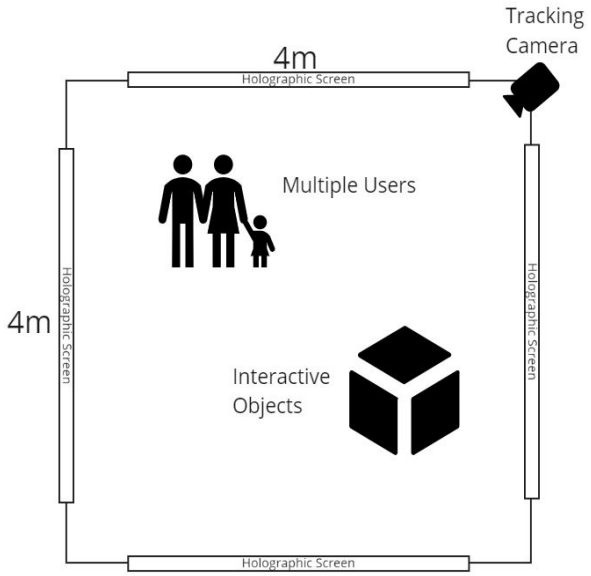


Figure 4.1: Prototype Floor Plan

and materials to build the frame using metal pipes and PVC connectors [1]. The necessary parts for each part of this prototype are listed as follows:

- 2 Horizontal Supports (arbitrary but identical length)
- 4 Vertical Supports (2m)
- 4 Short Length Weight Support Beams
- 2 L-Connectors, 4 V-Connectors, 2 3-Way Connectors
- Sufficient Length Holographic Screen

The width of the pipes was 2.5 m which, if doubled for each screen, meant that the frame would take up a 5x5 m^2 area. The screen itself, known as *Polid Screen* in Japan, is a greenhouse plastic sheet that has a height of 2 m, totaling the volume of the area to 5x5x2 m^3. The completed frame is photographed below.

Although the screens have a very wide dimension, they are only as large as the dimensions that the projector that illuminates them can support. This introduced

1 http://polidscreen.com/start.html

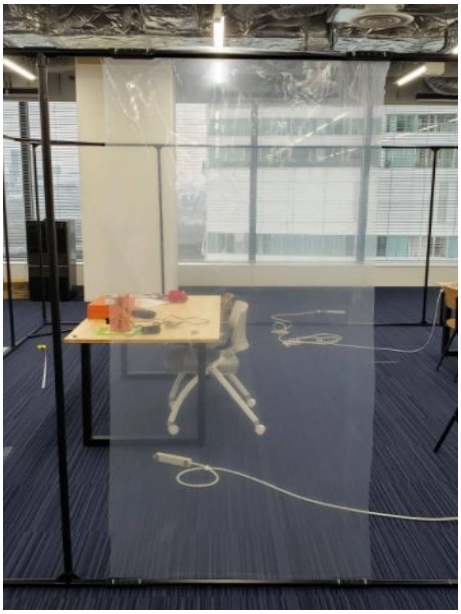


Figure 4.2: Thin Screen Prototype

Figure 4.3: Wide Screen Prototype

Figure 4.4: Wide Screen Attached to Frame

another variable into the setup, which was the ideal placement of each projector. Placing projectors behind the holographic screens required even more space; as regular cost-friendly projectors required a distance of 3m from the display to achieve a maximum height of 2m. This setup would keep the interior of the space empty, but would raise the area of the space absurdly, extending from 5m to 11m on each edge. There was also the concern of bleeding light from one display to the screen parallel to it, but through testing it was determined that the screens diminished light enough to not cause concern. This was likely due to the incorrect luminance level advertised on the projectors.

Since each projector required 3,000 lumens minimum to illuminate their respective screens, and available space was severely limited, the final design of the prototype made use of short throw projectors. Ultimately, short throw projectors are more expensive, but could be mounted to the ceiling of the space and directed at each display without concern for overlap. The benefit of using the short throw setup can be seen in Figure 4.5 where the user avoids projection blindness from within the space.

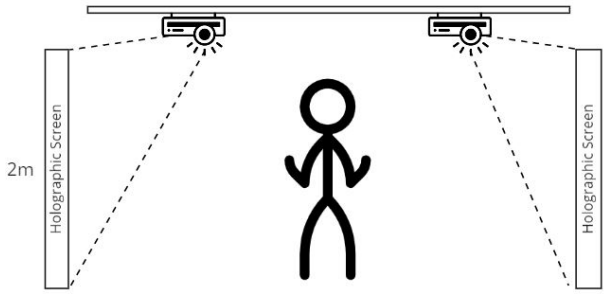


Figure 4.5: Short Throw Projector Diagram

Tracker Configuration

The next step was to equip the space with tracking sensors. In this prototype the most important sensors are the Microsoft Azure Kinect RGB-D Cameras, as they can capture point cloud data as well as body tracking data. The necessary components for the camera feed were as listed:

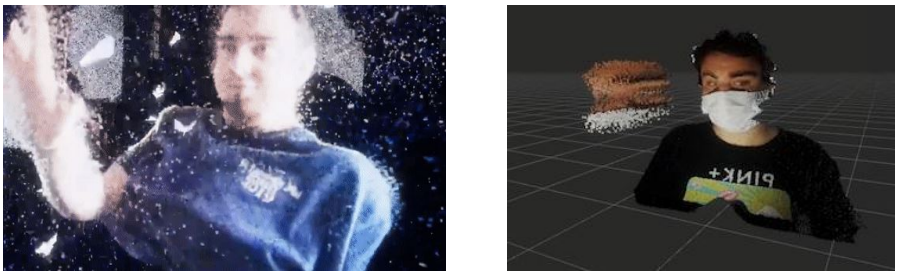
- Azure Kinect Cameras x2
- Tripod
- Ceiling Mount

To track a single user in the space is possible using one camera, therefore a single camera is mounted in one corner of the space. The other camera, which functions to funnel point cloud content into the virtual space, is not confined to either the inside or outside of the physical space. Therefore the second Kinect is simply attached to a tripod and is interchangeably mounted both inside and outside of the scene, depending on its purpose.

Point Clouds in Virtual Space

To perform data processing in this system, the following computer specifications are necessary, as well as the USB port requirements for the camera feed:

- NVidia RTX Equivalent Graphics Card
- 8GB RAM



a) Butterfly Space Distortion b) Facial Extrusion + Dissipation

Figure 4.6: Custom Virtual Effects

- USB 3.0 Ports x2
- Unity Game Engine
- 4 Output Display Ports

Everything on the digital side of this operation is handled in the Unity game engine - which serves as the programming interface for the virtual space. Whatever can be visualized in the Unity scene can then be reflected in the physical space, thus completing the feedback loop.

From a single Kinect, point cloud data is retrieved using a manager script. The manager script detects the camera, as well as its resolution, the near and far clipping boundaries, and assigns color values to one color map, while depth values are saved to a vertex map. Using a new feature introduced to Unity's High Definition Render Pipeline (HDRP) in 2019 known as the VFX graph, the vertex map and color map are used together to generate a point cloud which can be composed of millions of points. The shape, color, and size of the points can be configured using the VFX graph, to simulate lighting, noise, and virtual effects. After experimenting with the system for an extended period of time, some custom virtual effects I achieved are recorded in Figure 4.6.

The great thing about the VFX graph is that it is just a component on a game object, which means that it can be attached to another instance of an object, and the data can be duplicated in the same scene with little impact on the rendering speed. Each of the point clouds can then be treated as if the display they will exist on is a window in front of the data.

Body Tracking in Virtual Space

From another Kinect - using the same manager script - we can see body tracking information mapped to a skeleton object. This body can freely move around in virtual space synchronously with the physical space. Each part of the virtual skeleton is segmented and equipped with a collider which can allow for the skeleton to manipulate and bump into virtual objects even if they don't exist in the physical space. This is where the paradigm of virtual controller to virtual object interaction, stems from.

4.2.2 Programming the Localization

What I would consider the most significant contribution of this work is the localized point cloud interaction technique that I will describe in this section. To fully appreciate it's usability I will first address some major difficulties of working with RGB data.

Static data

In a static scene, where points are locked in a fixed location, it is technically possible to instantiate every point. Although the data set for a point cloud may be significantly large, containing the position and color of potentially millions of points, once loaded, interaction with those points using some type of virtual controller is feasible through colliders and rigid body components. Since doing this is incredibly computationally taxing there are other ways to render many points using a particle emission system.

Through a particle system, a small point cloud could be rendered with simulated physics, however these clouds would be severely limited by CPU limitations and are generally capped out in the range of thousands of points, significantly less than the millions that would compose a scene. Although possible, and severely computationally expensive, there is no ease on the system in the context of live data.

Live Data

Using the VFX graph in Unity has become the standard practice for handling live point cloud data. The VFX graph takes the position and color of all the input points, updates the data many times every second, and forwards it directly to the GPU to render the points on a display. This approach makes it possible to render millions of rapidly-updating particles from live data in a scene, and even to apply calculated effects using a node-based visual programming interface.

Limitations

What hasn't been widely explored, is the means through which we can interact with or change the display data in real time using a collision-capable virtual controller. The reason for this is that since the positional data comes from an unstructured map and is directly sent to the GPU, there is no collision detection among the points, which usually is performed by the CPU. Even worse, the spatial and color data of the point cloud is lost and irretrievable as it is sent through a one-way connection to the GPU.

It is because of this reason that many works featuring point cloud data will apply a pre-calculated effect to the entire scene on arbitrary intervals to simulate originality. But as a programmer, my goal is to create an interaction function that achieves varying levels of magnitude and visual influence based on the user's live input. A user can manage many variables in an interactive context such as spatial position, rotation, gestures, and can even perform slight calibration achieve a desired result.

Technical Contribution

An interactivity enabling node in the VFX graph is the *Kill (AA Box) node*. They simulate collision in a poor sense by taking basic geometric shapes and sending triggers in the rendering pipeline to eliminate particles within that shape's boundaries. After writing a script that would map the origin of a kill-box to a virtual controller, it was possible to erase sections of the scene manually, however the output of this interaction was rather boring. Since the points were killed without re-initialization, there was a severe lack of visual content.

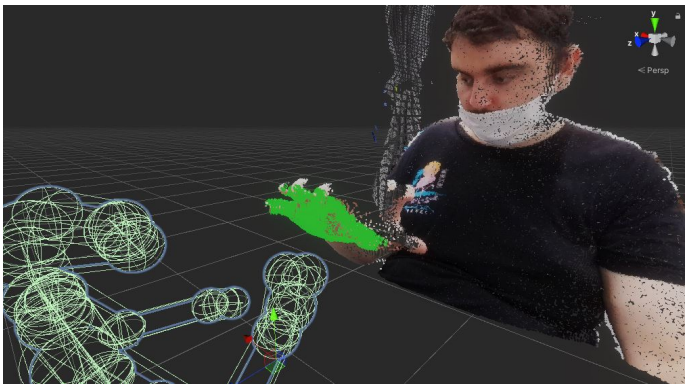
An experimental feature (still in beta development) of the VFX graph is the *collider box* which simulates collision by sending triggers through the rendering pipeline to translate particles outside of the boundaries of another geometric shape. By anchoring a collider box to a virtual controller via scripting, I was now able to temporarily warp particles within a radius of the controller to achieve a repelling interaction. This functionality provided highly responsive data-manipulation in localized areas, but altogether removed the possibility for any sense of "touch" or "approaching" certain points.

The next step was to make use of collider box inversion feature, which served better than kill boxes at only rendering points within a predefined geometric shape. Using this feature alone, a majority of the scene would vanish, except for those particles within the radius of the controller. As the controller moved in the virtual space, different sections of the scene would become visible. It was here where I was inspired to design a rendering technique that would enable localized effects in a live scene.

I created two VFX graphs, one that showed the general point cloud content, and a second *genuine effect graph* that would remain hidden except for within a radius of the virtual controller, as assigned in a script. The second graph could then be equipped with visual scripting patterns and effects that would be triggered upon proximity. To test this, using the body tracking API, the physical space user's hands, legs, and even nose were tracked by the system as the anchor for the effect graph. The result, after some calibration with the holographic screens, was a depth based interaction that circumvented object classification and scene segmentation that produced interesting localized effects. It also came with no significant drops in performance. The final result can be seen in Figure 4.7.

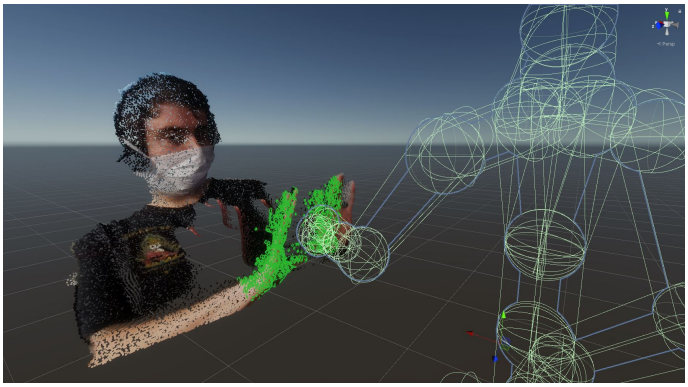
Directional Independence

Another positive feature of this design is that there is no cross-contamination between duplicated point clouds. In an omnidirectional setup, the same content can be displayed across multiple VFX graphs facing different directions, and only the graphs with data that are within radius of the controller in virtual space will be influenced by the *genuine effect graph*. This gives each instance of the point cloud data its own spatial uniqueness.



a) User approaching effect trigger radius of a mannequin

b) Precision effect interaction using a smaller radius

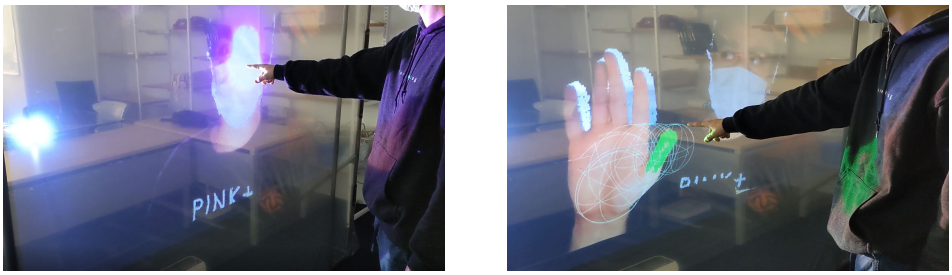


c) Physical user skeleton performing a left jab on virtual user

Figure 4.7: Technical Contribution: Depth Based Graphical Interaction

4.2.3 Functional Prototype

Multiple iterations of the prototype were built with varying display sizes as per the limitations of their respective locations. The final tangible output can be seen below (Figure 4.8) in the context of a user interacting with digital data of a remote user through virtual to virtual interaction. The system can be configured to also display the user's body tracking data on the holographic screen to assist with coordination and calibration in the virtual space (b).



a) Pink Burst Effect b) Body Tracking Reference

Figure 4.8: Virtual to Virtual Interaction

Since the demonstrated effects occurred on random time intervals, there was a reported sense of ambiguity with regards to what event triggered the virtual effect among beta testers. This feedback was positive for the system as it demonstrated that the interaction method disguised the fact that virtual effects were being triggered on computer generated random time intervals.

Another early feedback of this system that signaled concern for user well-being, was the strong light intensity. The first two iterations of this prototype were built before the strong intensity short throw projectors were obtained, and thus relied on two different varying brightness backlit projector systems. In the example above, the laser projector was sufficiently bright in a well-lit environment, but this made the intensity of certain high frequency projected colors (such as green and blue) too strong, causing user discomfort.

In the next figure (Fig 4.9), a different backlit prototype, using a projector of insufficient brightness is observed to cause *developer frustration* in an environment with moderate light pollution.

Figure 4.9: Sub-3000lm Laser Projection Setup

4.3. User Study

I believe that responsive bare-hand interaction with varying virtual objects would strongly impact the perceived quality of interaction and immersion based on recent egocentric interaction designs as outlined in Chapter 2. My hypothesis is that a user would therefore be inclined to interact with physical and virtual objects through the interfaces in a hybrid-space if the space is shown to be responsive. To test this hypothesis I designed an in-the-wild preliminary exhibition to measure the engagement and interest of users to validate the effectiveness and perceived quality of this toolkit.

4.3.1 Purpose

The purpose of the toolkit in this public exhibition is to serve as a framework that can support and facilitate the cross reality interaction. Thusly, the goal is to observe genuine user interest and curiosity in the environment, and to measure it's responsiveness and user friendliness in an *ideally* unsupervised hybrid-space. If users in the toolkit-supported-environment feel inclined to put their arms and legs out to cause an environmental reaction, and if the key interaction is proximity-based, then it is self-evident that bare-body distance based interaction scenarios can be desirable for public installations and merit further research.

Figure 4.10: Teddy-Bear Size Comparison with Water Bottle

4.3.2 Content

The preliminary exhibition was an observational experiment from a researcher's perspective, and was limited by restrictions of space, and equipment availability in the COVID-19 Pandemic. The setup in the exhibition consisted of 2 perpendicular adjacent holographic screens illuminated by 2 short throw projectors. On those screens, a simple and understandable content would be displayed that would compel users to observe the displays more closely. For this exhibition, a small teddy-bear was chosen for its age-friendliness and lightheartedness (Figure 4.10). To capture the teddy-bear model, as well as the data from the users in the environment, 2 RGB-D capable cameras (Microsoft Azure Kinect) were installed inside and outside of the environment for their unique roles.

The camera outside of the environment (Figure 4.11) would simply capture the visage of the teddy-bear and project it onto each of the displays, upscaled to nearly 1 meter tall, providing a larger presence within the scene (Figure 4.12). The camera inside the environment would track the body data of a single user, so that if they approached the display, a virtual effect would be generated (Figure 4.13). The two displays were configured with varying proximity thresholds, so that one interface could be used from a long range (1 meter away) while the other operated from close-proximity interaction (5-10 centimeters away).

The displayed teddy-bear model was also cloaked in a waving black ripple effect

Figure 4.11: Laptop Camera and Model Mount Setup

Figure 4.12: Projection Scale

Figure 4.13: Projection Interaction with Bare Hand

that made the image move subtly. This was intended to give the impression that since the content is moving, it must be live, and not just a static picture. The body probing performed in this scene was limited to only one user, but attached two probes to the user's left and right hands each applying a blue and pink recoloring effect respectively.

4.3.3 Procedure

The in-the-wild preliminary exhibition ran in a 3x3m square space for 2 days, each day operating for 3 hours. In that time, 20 participants (5 under the age of 10, 5 over the age of 40, 10 college-aged students) were gathered to try out the exhibition. Participants who were attending a local university event, were invited into a closed space to observe and freely interact with the environment. While users were present in the scene, a conductor would observe the behavior of participants and record if any of 3 intended interactions were achieved. The three objective interactions are as follows:

- The participant was able to calibrate themselves to the appropriate interactive distance of each display to achieve the color effect.

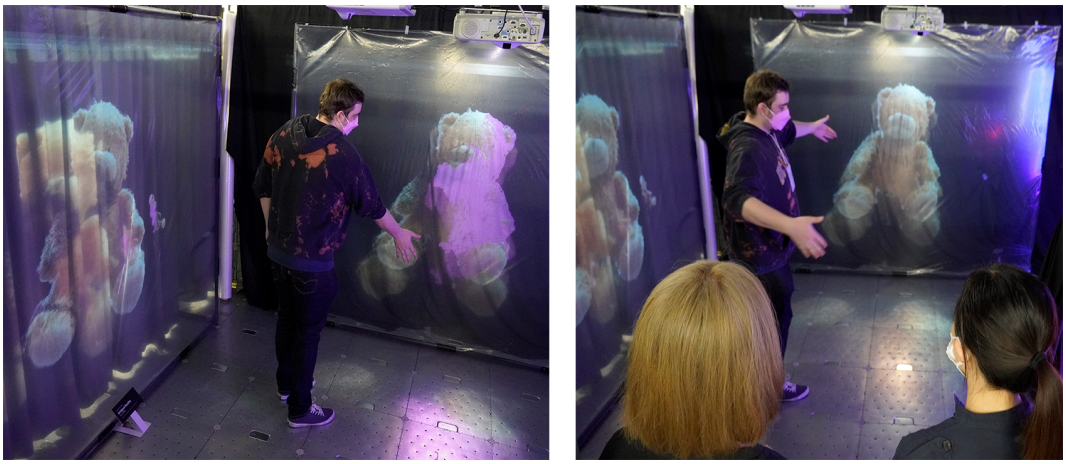


Figure 4.14: Conductor Demonstration from Inside Exhibit

- The participant discovered that using 2 different hands would produce 2 different colors, and attempted simultaneous interaction with both hands.
- The participant was inclined to attempt an "abnormal" interaction such as stretching to interact with both screens simultaneously, or using another part of their body.

Users who were confused by what they were seeing were offered an brief demonstration of the exhibit from a standby conductor (Figure 4.14). This was an opportunity to attempt objectives 2 and 3 which would be significantly more improbable if objective 1 was never accomplished before a user became disinterested. Participants were shown the optimal distance for each unique display to achieve color manipulation, and were then informed that the space is configured to best support 1 user at a time.

4.3.4 Results

Of the 20 participants observed who entered the exhibit, 14 (70%) had at least attempted, by the time they left, to interact with the holographic displays in some capacity. Since many users (11) accepted the offer for a conductor demonstration, it was made clear that the environment had a responsive quality to it. While not every user was effective at coordinating their movements with the system to

achieve a desired output, it can at least be seen that the hypothesis of this work was validated, that users were inclined to engage with an augmented environment if it could produce a response to the user.

Age-Related Findings

It was observed that 100% of participants age 10 and below instinctively approached the holographic interfaces and attempted to interact using their hands. Although inclination to walk up and touch the screens was also observed before the displays were even turned on. It was also observed that older participants were more inclined to not attempt interaction. Of the 5 eldest participants, 4 did not engage with the interface by moving their arms at all, and maintained a reserved, hands folded posture. One elder participant did express interest in the system and received a demonstration from the conductor, which then led to more intent driven activity from the participant, but often they did not notice that they were influencing the scene before their gaze was diverted.

Body-Tracking Challenges

Occasionally, users would enter the scene in groups which led to software-side system confusion, where only one participant would be equipped with the effect creating probes, and no way to tell who it was. It was also noted that the camera placement in an upper corner of the scene, caused difficulties in virtual space calibration, and user tracking. This could potentially be improved in future implementations by installing the camera closer to the ground or even potentially behind the holographic screens due to their inherent transparency.

Imaginative Interaction

Participants who learned the optimal distances for interaction spent longer in the scene trying to achieve desired effects than those who quit before acclimating to the distance. 3 participants attempted to interact with both screens simultaneously by stretching their arms to match the proximity difference between interfaces. Another participant was so engaged in the system that they requested

the probe to be shifted to their feet, so that they could "kick" the virtual data, which was easily configurable by our toolkit.

4 participants in the study were interested enough in the technical setup of the system that they were invited behind-the-scenes by a conductor to view the toolkit setup. All 4 participants expressed shock to find out that the exhibition was running on a laptop, and that the teddy-bear in the scene was live data. Two participants replaced the teddy-bear with themselves and interacted with each other's live point cloud data using the environment in a teleconferencing style.

Lastly, the addition of user's point clouded hands or face on the holographic screens was not always immediately noticed. Similarly, many observed instances of users causing color shifted interactions in the scene went unnoticed by participants who were observed to be less interested.

4.3.5 Discussion

From our early results, it was seen that younger audiences were more inclined to interact with the exhibit, while older audiences behaved in with more reservation. Therefore, it is believed that initial engagement is socially influenced, or a learned behavior. This means that with the popularization of depth based RGB-D interaction toolkits such as *Bridged Reality*, expectations of environmental responsiveness may be learned. This would result in a more common desire to interact with holographic displays.

This work proposes that there should exist an optimal ratio of display scale to user FOV, derived from proximity, making produced interactions more visually apparent. This would help users who may not notice the area of the display that they are influencing, become more aware of their presence in the system. A dynamic display that adjusts its visual projection depending on where it's observer is standing is also an interesting application that could be programmed in a future implementation of this toolkit.

In terms of setup simplicity, the toolkit was pre-configured with a variety of visual effect graphs, several scripts that handled the configuration of probing. As a machine that has many turning gears, designing a hybrid-space from scratch can be intimidating, exhausting, and require in-depth programming knowledge. This toolkit addressed that barrier to customization by packaging effects, and config-

uration into a short process, that was demonstrated to be understandable and configurable by exhibition participants, as well as a technically inclined middle-schooler. A tutorial with a screenshot of an example VFX graph can be found in Appendix A.

For facilitating interaction, it was shown that a single camera was sufficient to perform bare-body user tracking for this interaction design; a cost effective alternative to virtual controllers and wearable devices for open-area interfaces. However, it is important to measure user comfortability and satisfaction with this method of tracking, which was not collected in this preliminary exhibition. Hence, a user feedback survey will be deployed in the future to gauge system usability via System Usability Scale, and user enjoyment via Self-Assessment Mankikin Scale and a Frustration Discomfort Scale.

4.4. Measurement of Usefulness

One of the necessary usefulness checkpoints of this thesis is for the proposed toolkit to be demonstrably effective in facilitating the types of interaction that the system designer wishes to take place. I will outline in this section the future methods of data collection for this prototype, and their implications in the conclusion of this work.

As for data collection, there will be two primary types of feedback that measure the system's effectiveness by use. One will be in the form of an optional survey that will be composed of questions from three well-established evaluation metrics. To avoid the psychological disconnect between users and their experience with augmented reality systems through intense surveys, the survey will be composed of 10 short questions specific to the expected interactive content. The questions will include 4 System Usability Scale (SUS) focused queries, 3 Self-Assessment Manikin (MAN) scale queries, and 3 Frustration Discomfort Scale (FDS). These three categories were deemed positive indicators of whether users feel that they are achieving the kind of interaction they desire, if the system flags any immediate kinds of discomfort, and primarily the system has market potential and general usability.

The second form of feedback should be through observation from the perspec-

tive of the designer. In a future user study, there will be several *hidden* custom interactions that will be accessible to all users, but likely not achieved by all. By monitoring users to determine if the hidden interaction was observed, this feedback will provide critical information about the UX design that would trigger such an event. The feedback would clarify predictable actions, and unlikely gestures in the current hybrid-space.

The hidden interactions can include scenarios from the preliminary exhibition such as a "multiple hand enabled effect", or new designs such as a "secret section of the display", or "an input data replacement". Briefly put, these activities will signal to the designer whether a user is intuitively trying to use two hands for an interaction, if they attempt to interact with enough of the display to find an abnormal section, or if they are inclined to change the input data of the system using an in-scene RGB-D Camera. This feedback is critical for designing interaction scenarios that surpass the need for language and explicit instruction.

4.4.1 How Responsiveness Contributes to Incentivization

This list addresses the relationship between the survey results and toolkit application:

- Positive SUS : Potential Marketability
- Negative SUS : Unpopular System Design
- Positive MAN : Effective User Control
- Negative MAN : Ineffective User Control
- Positive FDS : Satisfactory User Comfort
- Negative FDS : Ethical Concerns & Safety Limitations

From the secondary feedback of hidden task completion, the data will be treated as a measurement of whether or not users felt inclined to test the system enough to generate a new experience. If the feedback of this data shows that a significant number of users discovered the hidden interactions, then it can be inferred that the toolkit was used in a compelling way from the designer's perspective.

4.4.2 Social Impact of Cross-reality Systems

As is typical for nearly all reality mixing works, the application and social impact of this toolkit is highly derivative of the user feedback from demoing the system. Alpha stage testing of this toolkit has produced a new interaction technique with particle data that has not been seen in recent literature, and until recently, would not have been possible with such high responsiveness.

Let us then center the social impact around this interactivity alone. In the context of casual gaming, this system provides an opportunity for game developers to augment high definition, dense graphical renderings around a depth based input from the user through a camera. Since effect-based cosmetics are infamously desirable across many genres, an interactive effect that takes data from the user's physical space would be quite interesting to see in multiplayer scenarios. In the near future, a gamer could potentially import a point cloud of their dog into a game as a virtual companion, or control a character's hands and gesture with their own bare-hands.

In the workplace environment, the toolkit can also offer an exciting new experience for people who work with data visualization. For example any data set (such as stock prices, or population density) that includes points and some given color representation, could be projected into 3D space and manipulated by hand. This new interactive scenario can lead analysts away from typical virtual controllers like a mouse and keyboard to more abstract and engaging methods of input made possible through a camera.

Another social impact of the multi directional design of this work can be demonstrated in a collaborative work environment. For example a conference call could be segmented across displays so that one user can be more immersed by looking in different directions at different speakers. The user would also have the freedom to interact with the visual data of the conference attendees and perform various data-distortion activities on the camera input such as drawing, erasing, voxalizing, or recoloring the feed.

Finally, as the main theme of this thesis, this tool kit can give new purpose to empty space through augmentation, as well as more functionality to in-use space, via a virtual counterpart. Perhaps a user may have no more space in their bedroom for sentimental things like photos or stuffed animals. But through *Bridged Reality*

the user can copy those things into the virtual space which could then still be dynamically rendered back into the physical space; effectively providing more storage room in the physical space.

Chapter 5
Conclusion

5.1. Summary, Future Works and Limitations

This thesis summarized existing applications of point cloud data in Chapter 1. It also brought to light some of the limitations of physical spaces, primarily focusing on their lack of interactivity to some user groups. A proposed interaction feedback loop, taking place across physical and virtual realities, was offered as a means to give purpose to boring space. A target group of users, from lower economic background and computer unfamiliarity, was also identified as the key audience for this work.

Chapter 2 evaluated a variety of relevant works across a broad spectrum of digital media technology, and accented desirable features in the context of digital spatial augmentation and interaction. It was found that point cloud data had much potential to function as a facilitator for virtual interactions, and that light field projection techniques produced highly immersive interaction scenarios.

Chapter 3 then outlined the design of a proposed toolkit aimed to encourage the use and design of RGB-D data driven interactions. The toolkit would incorporate the aforementioned desirable features into a simple package that could be easily understood and operated by the less tech savvy, or even programmers that simply want to introduce a new interaction medium to their system.

In Chapter 4 the blueprint and implementation of a prototype is described. Multiple iterations of the prototype were built, so their condensed strengths and demonstrated content are documented. An in-depth programming journey that resulted in an fast-acting, classification-avoiding, rendering pipeline manipulation technique is also shared. A preliminary exhibition that serves as the user study for this thesis is outlined. Exhibition results and some initial feedback of the functioning system is condensed, leading to the proposal a formal observational

and survey-based evaluation metric of the toolkit. Finally, the social impact of the toolkit and it's applications are projected.

In conclusion, it is this author's opinion that the proposed input and data processing pipeline - as outlined in Chapter 3, and demonstrated in Chapter 4 - achieves an uncommon and unique depth based interaction. While avoiding the complexities of deep learning algorithms and overly complicated data-set analysis, a highly responsive and visually aesthetic output is rendered using game design techniques. Moreover the output holographic display medium of the system, while interesting in nature and highly anticipated, added significant cost to the system.

Limitations

Due to limitations with available space during this research process, ambient light pollution also heavily reduced the visibility of some prototypes. It is therefore an inherent limitation of this work to be operated in a dark space if the output medium is a projection holographic display. Another limitation of the work is the need for a specific RGB-D camera, namely the Microsoft Azure Kinect - which has been out of stock in the COVID-19 Pandemic - as a result of the manager scripts which are only compatible with this device.

Future Work

The interactive effects observed through the use of this toolkit are applicable in a variety of applications due to the nature of common human-object interaction scenarios as outlined in Chapter 3. Through a more complex physics engine, derivative depth-based interactions such as elasticity, kinetics, and deformative manipulations can and should be explored, as they were not seen in the scope of this prototype. The resulting visual effects that were demonstrated in Chapter 4 however, are implementation ready and can be used in applications such as game design, art installations, environmental re-construction, and even high precision agricultural observation as mentioned in Chapter 1.

References

[1] Ivan E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, AFIPS '68 (Fall, part I), pages 757–764, New York, NY, USA, December 1968. Association for Computing Machinery. URL: http://doi.org/10.1145/1476589.1476686, doi:10.1145/1476589.1476686.

[2] Preeti Sirohi, Amit Agarwal, and Piyush Maheshwari. A survey on Augmented Virtual Reality: Applications and Future Directions. In *2020 Seventh International Conference on Information Technology Trends (ITT)*, pages 99–106, November 2020. doi:10.1109/ITT51279.2020.9320869.

[3] Mark Graham, Matthew Zook, and Andrew Boulton. Augmented reality in urban places: contested content and the duplicity of code: *Augmented reality in urban places. Transactions of the Institute of British Geographers*, 38(3):464–479, July 2013. URL: http://doi.wiley.com/10.1111/j.1475-5661.2012.00539.x, doi:10.1111/j.1475-5661.2012.00539.x.

[4] Louis Rosenberg. The Use of Virtual Fixtures as Perceptual Overlays to Enhance Operator Performance in Remote Environments. page 52, September 1992.

[5] Frank J. Delgado, Michael F. Abernathy, Janis White, and William H. Lowrey. Real-time 3D flight guidance with terrain for the X-38. In Jacques G. Verly, editor, *Enhanced and Synthetic Vision 1999*, volume 3691, pages 149 – 156. International Society for Optics and Photonics, SPIE, 1999. URL: https://doi.org/10.1117/12.354416, doi:10.1117/12.354416.

[6] Augmented Reality Is Finally Getting Real. URL: https://www.technologyreview.com/2012/08/02/184660/augmented-reality-is-finally-getting-real/.

[7] Jannick Rolland, Frank Biocca, Felix Hamza-Lup, and Yanggang Ha. Development of head-mounted projection displays for distributed, collaborative, augmented reality applications. *Presence-Teleoperators and Virtual Environments*, 14(5):528–549, January 2005. URL: https://stars.library.ucf.edu/facultybib2000/5607, doi:10.1162/105474605774918741.

[8] Leonel Merino, Magdalena Schwarzl, Matthias Kraus, Michael Sedlmair, Dieter Schmalstieg, and Daniel Weiskopf. Evaluating Mixed and Augmented Reality: A Systematic Literature Review (2009-2019). In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 438–451, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00069.

[9] Sukirman, Ika Fitri Nur Janah, Reza Arif Wibisono, and Nur Subekti. Visualizing 3D Objects Using Augmented Reality Application to Enhance Students Retention in Social Science Subject. In *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 127–132, September 2019. doi:10.1109/ISEMANTIC.2019.8884318.

[10] Gabriel Freitas, Marcio Sarroglia Pinho, Milene Selbach Silveira, and Frank Maurer. A Systematic Review of Rapid Prototyping Tools for Augmented Reality. In *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, pages 199–209, November 2020. doi:10.1109/SVR51698.2020.00041.

[11] Saiwen Wang, Jie Song, Jaime Lien, Ivan Poupyrev, and Otmar Hilliges. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 851–860, New York, NY, USA, October 2016. Association for Computing Machinery. URL: http://doi.org/10.1145/2984511.2984565, doi:10.1145/2984511.2984565.

[12] Bonchang Koo, Joonho Kim, and Jundong Cho. Leap motion gesture based interface for learning environment by using leap motion. In *Proceedings of HCI Korea*, HCIK '15, pages 209–214, Seoul, KOR, December 2014. Hanbit Media, Inc.

[13] Richard Sahala Hartanto, Ryoichi Ishikawa, Menandro Roxas, and Takeshi Oishi. A Hand Motion-guided Articulation and Segmentation Estimation. *arXiv:2005.03691 [cs]*, May 2020. URL: http://arxiv.org/abs/2005.03691.

[14] Gregory Izatt, Geronimo Mirano, Edward Adelson, and Russ Tedrake. Tracking objects with point clouds from vision and touch. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4000–4007, Singapore, Singapore, May 2017. IEEE. URL: http://ieeexplore.ieee.org/document/7989460/, doi:10.1109/ICRA.2017.7989460.

[15] Richard A Newcombe, Andrew J Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. KinectFusion: Real-Time Dense Surface Mapping and Tracking. page 10.

[16] Simone Teruggi, Eleonora Grilli, Michele Russo, Francesco Fassi, and Fabio Remondino. A Hierarchical Machine Learning Approach for Multi-Level and Multi-Resolution 3D Point Cloud Classification. *Remote Sensing*, 12(16):2598, August 2020. URL: https://www.mdpi.com/2072-4292/12/16/2598, doi:10.3390/rs12162598.

[17] Mikaela Angelina Uy, Jingwei Huang, Minhyuk Sung, Tolga Birdal, and Leonidas Guibas. Deformation-Aware 3D Model Embedding and Retrieval. *arXiv:2004.01228 [cs, eess]*, July 2020. URL: http://arxiv.org/abs/2004.01228.

[18] Antoine Petit, Stephane Cotin, Vincenzo Lippiello, and Bruno Siciliano. Capturing Deformations of Interacting Non-rigid Objects Using RGB-D Data. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 491–497, Madrid, October 2018. IEEE. URL: https://ieeexplore.ieee.org/document/8593756/, doi:10.1109/IROS.2018.8593756.

[19] Martin Simon, Karl Amende, Andrea Kraus, Jens Honer, Timo Sämann, Hauke Kaulbersch, Stefan Milz, and Horst Michael Gross. Complexer-YOLO:

Real-Time 3D Object Detection and Tracking on Semantic Point Clouds. *arXiv:1904.07537 [cs]*, April 2019. URL: http://arxiv.org/abs/1904.07537.

[20] S. Wang, Y. Sun, C. Liu, and M. Liu. PointTrackNet: An End-to-End Network For 3-D Object Detection and Tracking From Point Clouds. *IEEE Robotics and Automation Letters*, 5(2):3206–3212, April 2020. doi:10.1109/LRA.2020.2974392.

[21] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. ZoomNet: Part-Aware Adaptive Zooming Neural Network for 3D Object Detection. *arXiv:2003.00529 [cs]*, March 2020. URL: http://arxiv.org/abs/2003.00529.

[22] Demetri Terzopoulos, Andrew Witkin, and Michael Kass. Constraints on deformable models:Recovering 3D shape and nonrigid motion. *Artificial Intelligence*, 36(1):91–123, August 1988. URL: http://www.sciencedirect.com/science/article/pii/000437028890080X, doi:10.1016/0004-3702(88)90080-X.

[23] Lihui Wang and Baozong Yuan. Curvature and density based feature point detection for point cloud data. In *IET 3rd International Conference on Wireless, Mobile and Multimedia Networks (ICWMNN 2010)*, pages 377–380, September 2010. doi:10.1049/cp.2010.0694.

[24] Yang You, Yujing Lou, Qi Liu, Yu-Wing Tai, Lizhuang Ma, Cewu Lu, and Weiming Wang. Pointwise Rotation-Invariant Network with Adaptive Sampling and 3D Spherical Voxel Convolution. *arXiv:1811.09361 [cs]*, December 2019. URL: http://arxiv.org/abs/1811.09361.

[25] Qingdong He, Zhengning Wang, Hao Zeng, Yi Zeng, Shuaicheng Liu, and Bing Zeng. SVGA-Net: Sparse Voxel-Graph Attention Network for 3D Object Detection from Point Clouds. *arXiv:2006.04043 [cs]*, June 2020. URL: http://arxiv.org/abs/2006.04043.

[26] Y. Lee and S. Seo. Real-Time Object Tracking in Sparse Point Clouds Based on 3D Interpolation. In *2018 IEEE International Conference on Robotics*

and Automation (ICRA), pages 4804–4811, May 2018. doi:10.1109/ICRA.2018.8460639.

[27] Winston H. Hsu. Learning from 3D (Point Cloud) Data. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2697–2698, New York, NY, USA, October 2019. Association for Computing Machinery. URL: http://doi.org/10.1145/3343031.3350540, doi:10.1145/3343031.3350540.

[28] Qiangeng Xu, Xudong Sun, Cho-Ying Wu, Panqu Wang, and Ulrich Neumann. Grid-GCN for Fast and Scalable Point Cloud Learning. page 10.

[29] D. Li, H. Wang, N. Liu, X. Wang, and J. Xu. 3D Object Recognition and Pose Estimation From Point Cloud Using Stably Observed Point Pair Feature. *IEEE Access*, 8:44335–44345, 2020. doi:10.1109/ACCESS.2020.2978255.

[30] Kazuma Yoshino, Hiroyuki Kawakita, Takuya Handa, and Kensuke Hisatomi. Viewing Style of Augmented Reality/Virtual Reality Broadcast Contents while Sharing a Virtual Experience. In *26th ACM Symposium on Virtual Reality Software and Technology*, pages 1–3, Virtual Event Canada, November 2020. ACM. URL: https://dl.acm.org/doi/10.1145/3385956.3422110, doi:10.1145/3385956.3422110.

[31] David Dunn, Okan Tursun, Hyeonseung Yu, Piotr Didyk, Karol Myszkowski, and Henry Fuchs. Stimulating the Human Visual System Beyond Real World Performance in Future Augmented Reality Displays. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 90–100, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00029.

[32] Etienne Peillard, Yuta Itoh, Guillaume Moreau, Jean-Marie Normand, Anatole Lécuyer, and Ferran Argelaguet. Can Retinal Projection Displays Improve Spatial Perception in Augmented Reality? In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 80–89, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00028.

[33] Olaf Kahler, Victor Adrian Prisacariu, Carl Yuheng Ren, Xin Sun, Philip Torr, and David Murray. Very High Frame Rate Volumetric Integration of

Depth Images on Mobile Devices. *IEEE Trans. Visual. Comput. Graphics*, 21(11):1241–1250, November 2015. URL: https://ieeexplore.ieee.org/document/7165673/, doi:10.1109/TVCG.2015.2459891.

[34] Songhai Zhang, Xiangli Li, Yingtian Liu, and Hongbo Fu. Scale-aware Insertion of Virtual Objects in Monocular Videos. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 36–44, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00022.

[35] D. Nam, J. Lee, Y. H. Cho, Y. J. Jeong, H. Hwang, and D. S. Park. Flat Panel Light-Field 3-D Display: Concept, Design, Rendering, and Calibration. *Proceedings of the IEEE*, 105(5):876–891, May 2017. doi:10.1109/JPROC.2017.2686445.

[36] T. Agocs, T. Balogh, T. Forgacs, F. Bettio, E. Gobbetti, G. Zanetti, and E. Bouvier. A Large Scale Interactive Holographic Display. In *IEEE Virtual Reality Conference (VR 2006)*, pages 311–311, Alexandria, VA, USA, 2006. IEEE. URL: http://ieeexplore.ieee.org/document/1667676/, doi:10.1109/VR.2006.9.

[37] Lixia Ni, Zhenxing Li, Haifeng Li, and Xu Liu. 360-degree large-scale multiprojection light-field 3D display system. *Appl. Opt., AO*, 57(8):1817–1823, March 2018. URL: https://www.osapublishing.org/ao/abstract.cfm?uri=ao-57-8-1817, doi:10.1364/AO.57.001817.

[38] Mohammad Hosseini and Christian Timmerer. Dynamic Adaptive Point Cloud Streaming. In *Proceedings of the 23rd Packet Video Workshop*, PV '18, pages 25–30, New York, NY, USA, June 2018. Association for Computing Machinery. URL: http://doi.org/10.1145/3210424.3210429, doi:10.1145/3210424.3210429.

[39] T. Yoshida, K. Shimizu, T. Kurogi, S. Kamuro, K. Minamizawa, H. Nii, and S. Tachi. RePro3D: full-parallax 3D display with haptic feedback using retro-reflective projection technology. In *2011 IEEE International Symposium on VR Innovation*, pages 49–54, March 2011. doi:10.1109/ISVRI.2011.5759601.

[40] HaiyangYu, XiaoyuJiang, XingpengYan, Zhiqiang Yan, Chenqing Wang, Zhan Yan, and Fenghao Wang. Research Summary on Light Field Display Technology Based on Projection. *J. Phys.: Conf. Ser.*, 1682:012033, November 2020. URL: https://doi.org/10.1088/1742-6596/1682/1/012033, doi:10.1088/1742-6596/1682/1/012033.

[41] Woojin Cho, Gabyong Park, and Woontack Woo. Bare-hand Depth Inpainting for 3D Tracking of Hand Interacting with Object. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 251–259, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00048.

[42] Etienne Dubeau, Mathieu Garon, Benoit Debaque, Raoul de Charette, and Jean-François Lalonde. RGB-D-E: Event Camera Calibration for Fast 6-DOF object Tracking. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 127–135, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00034.

[43] James Coleman Eubanks, Alec G. Moore, Paul A. Fishwick, and Ryan P. McMahan. The Effects of Body Tracking Fidelity on Embodiment of an Inverse-Kinematic Avatar for Male Participants. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 54–63, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00025.

[44] Alejandro Martin-Gomez, Alexander Winkler, Kevin Yu, Daniel Roth, Ulrich Eck, and Nassir Navab. Augmented Mirrors. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 217–226, November 2020. ISSN: 1554-7868. doi:10.1109/ISMAR50242.2020.00045.

[45] M. Billinghurst and T. Starner. Wearable devices: new ways to manage information. *Computer*, 32(1):57–64, January 1999. URL: http://ieeexplore.ieee.org/document/738305/, doi:10.1109/2.738305.

[46] Tibor Balogh, Zsuzsa Dobranyi, Tamas Forgacs, Attila Molnar, Laszlo Szloboda, Enrico Gobbetti, Fabio Marton, Fabio Bettio, Gianni Pintore, Gianluigi Zanetti, Eric Bouvier, and Reinhard Klein. An interactive multi-user holographic environment. In *ACM SIGGRAPH 2006 Emerging technologies*,

SIGGRAPH '06, pages 18–es, New York, NY, USA, July 2006. Association for Computing Machinery. URL: http://doi.org/10.1145/1179133.1179152, doi:10.1145/1179133.1179152.

[47] Mikito Takenaka, Takashi Kakue, Tomoyoshi Shimobaba, and Tomoyoshi Ito. Interactive Holographic Display for Real-Time Drawing and Erasing of 3D Point-Cloud Images With a Fingertip. *IEEE Access*, 9:36766–36774, 2021. Conference Name: IEEE Access. doi:10.1109/ACCESS.2021.3062877.

[48] Nicolas Villar, Daniel Cletheroe, Greg Saul, Christian Holz, Tim Regan, Oscar Salandin, Misha Sra, Hui-Shyong Yeo, William Field, and Haiyan Zhang. Project Zanzibar: A Portable and Flexible Tangible Interaction Platform. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery. URL: http://doi.org/10.1145/3173574.3174089, doi:10.1145/3173574.3174089.

[49] Xiao Tang, Xiaowei Hu, Chi-Wing Fu, and Daniel Cohen-Or. GrabAR: Occlusion-aware Grabbing Virtual Objects in AR. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pages 697–708, New York, NY, USA, October 2020. Association for Computing Machinery. URL: https://doi.org/10.1145/3379337.3415835, doi:10.1145/3379337.3415835.

[50] David Lindlbauer and Andy D. Wilson. Remixed Reality: Manipulating Space and Time in Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA, April 2018. Association for Computing Machinery. URL: https://doi.org/10.1145/3173574.3173703, doi:10.1145/3173574.3173703.

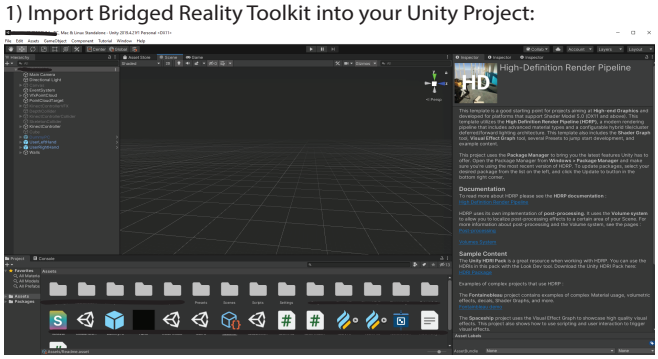
[51] Brett R. Jones, Hrvoje Benko, Eyal Ofek, and Andrew D. Wilson. IllumiRoom: immersive experiences beyond the TV screen. *Communications of the ACM*, 58(6):93–100, May 2015. URL: http://doi.org/10.1145/2754391, doi:10.1145/2754391.

Appendices

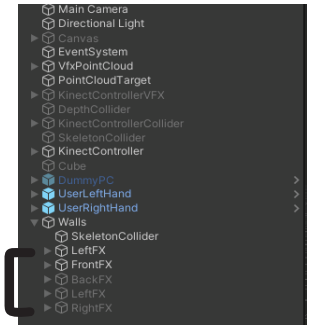
A. Bridged Reality Configuration Document

BRIDGED REALITY EFFECT CONFIGURATION TUTORIAL

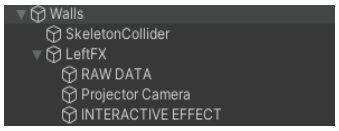
1) Import Bridged Reality Toolkit into your Unity Project:



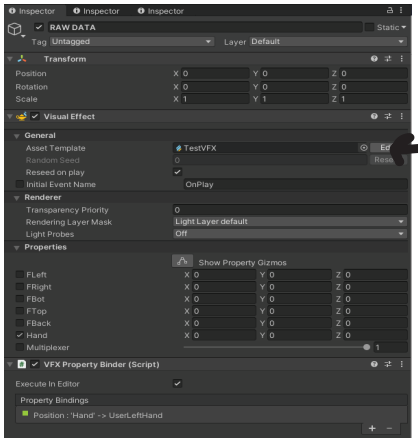
2) Select the direction of the display you want to configure



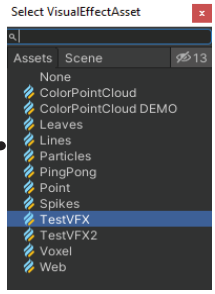
3) Select if you want the data, or the interaction effect to be changed.

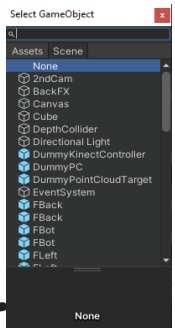


4) In the inspector tab, choose the VFX asset template you want to apply



5) Play the scene

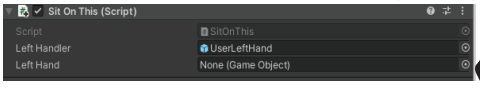




PROBE CONFIGURATION TUTORIAL

1) Play the scene
2) Select UserLeftHand or UserRightHand

3) In the inspector tab, find the Sit On This script component.
4) Change the hand, to any object in the scene. For body probing, Search "HandLeft" or "FootRight" or "Nose" for varying body parts.



Feel free to try to edit & configure your own VFX GRAPHS!