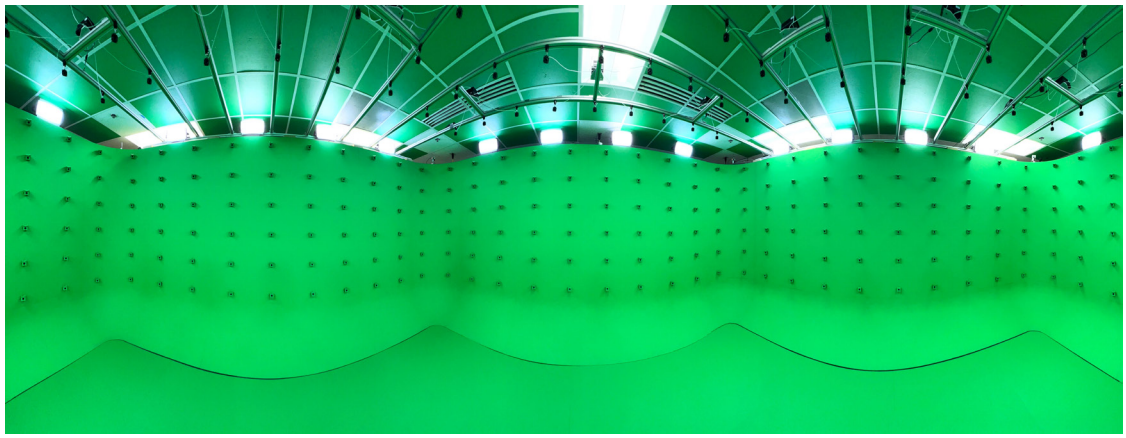
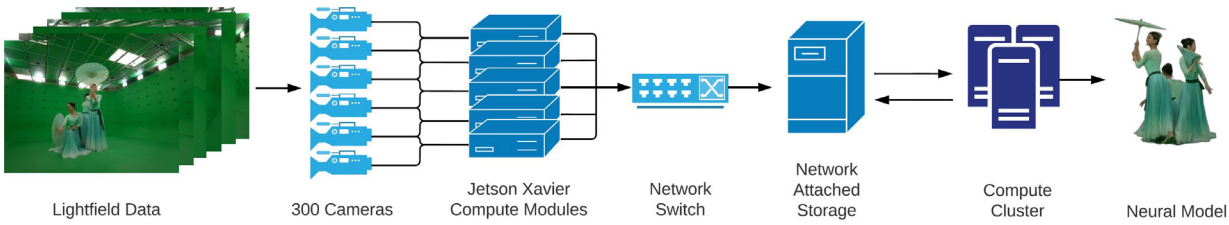
HoloCamera: Advanced Volumetric Capture for Cinematic-Quality VR Applications

Jonathan Heagerty*, Sida Li*, Eric Lee, Shuvra Bhattacharyya, Sujal Bista, Barbara Brawn, Brandon Y. Feng, Susmija Jabbireddy, Joseph JaJa, Hernisa Kacorri, David Li, Derek Yarnell, Matthias Zwicker, and Amitabh Varshney



(a) Panoramic picture of HoloCamera showing the infrastructure with 300 cameras mounted on the sides and ceiling of the capture studio.



(b) Overview of our system workflow. We capture lightfield data through a meticulously designed camera, network, storage, compute system and process the data into a volumetric neural model.

Abstract— High-precision virtual environments are increasingly important for various education, simulation, training, performance, and entertainment applications. We present HoloCamera, an innovative volumetric capture instrument to rapidly acquire, process, and create cinematic-quality virtual avatars and scenarios. The HoloCamera consists of a custom-designed free-standing structure with 300 high-resolution RGB cameras mounted with uniform spacing spanning the four sides and the ceiling of a room-sized studio. The light field acquired from these cameras is streamed through a distributed array of GPUs that interleave the processing and transmission of 4K resolution images. The distributed compute infrastructure that powers these RGB cameras consists of 50 Jetson AGX Xavier boards, with each processing unit dedicated to driving and processing imagery from six cameras. A high-speed Gigabit Ethernet network fabric seamlessly interconnects all computing boards. In this systems paper, we provide an in-depth description of the steps involved and lessons learned in constructing such a cutting-edge volumetric capture facility that can be generalized to other such facilities. We delve into the techniques employed to achieve precise frame synchronization and spatial calibration of cameras, careful determination of angled camera mounts, image processing from the camera sensors, and the need for a resilient and robust network infrastructure. To advance the field of volumetric capture, we are releasing a high-fidelity static light-field dataset, which will serve as a benchmark for further research and applications of cinematic-quality volumetric light fields.

Index Terms—Volumetric Capture, Light Fields, Holoportation, Multi-camera Array

1 Introduction

While there are several options to create digital humans, current technology solutions, primarily driven by hand-crafted art or 2D video, lack critical human factors necessary for experiencing true telepresence. These factors include eye contact, facial micro-expressions, natural body language, person-to-person engagement, and effective participation. Recent technological advances have resulted in digital cameras that can now exceed human visual acuity. However, they can only reproduce reality faithfully from one viewpoint. The moment a user's viewpoint changes from the point of view of the camera image, which is necessary for interacting with immersive virtual environments, the previously captured camera image can no longer faithfully reproduce

- *Correspondence to: jheager2@umd.edu, sidali@umd.edu*
- *All authors are with the University of Maryland, College Park.*
- **Joint first authors with equal contribution.*

Manuscript received 4 October 2023; revised 17 January 2024; accepted 24 January 2024. Date of publication 2 April 2024; date of current version 15 April 2024. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3372123, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2024.3372123

reality. That is why, even though high-resolution videos are used for safety-critical applications such as flight control, users do not feel comfortable substituting face-to-face interactions with current-generation telepresence technologies. High-fidelity digital human avatars and their interaction environments are critical for increasing communication, empathy, trust, critical thinking, and decision-making in immersive education, training, and entertainment.

Three decades ago, we witnessed a significant and swift transition in the imaging field from film to digital technology. This transformation was made possible by rapid advancements in photonic-sensor technology, the development of high-resolution displays, and improvements in computational capabilities for storing and processing digital images with print-quality clarity. We now find ourselves in the initial phases of the next major shift in imaging. This shift revolves around acquiring, processing, and presenting digital light fields, encompassing the collection of light rays that capture a scene from multiple perspectives. Three key factors drive this emerging paradigm:

- Rapid advancements in computational photography have led to novel methods that efficiently capture, interpolate and organize extensive light field data into coherent data structures.
- Efficient machine-learning algorithms well-suited to the latest highly parallel streaming processor architectures facilitate the rapid processing of these intricate light fields.
- Swift progress in consumer virtual reality (VR) displays, including holographic displays, which are poised to showcase the acquired light fields directly, revolutionizing the way we perceive and interact with visual content.

This paper presents the challenges encountered and lessons learned in building a high-quality volumetric capture facility. Specifically, we have designed and developed the HoloCamera, a cutting-edge facility comprised of a unique, spatially enveloping array of cameras and a computational cluster to facilitate the acquisition, processing, and generation of precision virtual environments for scientific exploration, knowledge discovery, training, and education. This has required developing new end-to-end advances in volumetric capture systems, interleaving acquisition with processing, and scaling up to an extensive collection of calibrated ensembles of cameras. The HoloCamera and its resulting light-field dataset, which we plan to release with the publication of this paper, will facilitate transformative research in several fields, including graphics, virtual environments, generative AI, high-performance computing, and signal processing.

2 Related Works

Several capture systems have been developed for capturing light fields and volumetric data. We provide a brief overview of some such systems.

2.1 Academic Research Studios

The Stanford Multi-Camera Array is a well-known camera system, used in various volumetric and light field rendering research projects since it was first introduced in 2002 [37, 38]. It consisted of 100 cameras capable of capturing 30 frames per second at a resolution of 640 $\times$ 480. The videos were synchronized to 0.1 milliseconds using the broadcast IEEE1394 command. This camera array employed several different setups, each designed to serve a specific purpose. In the first setup, the cameras were arranged in a plane with most of their fields of view overlapping at a distance of 10 feet. The second setup involved staggering the trigger times to create a virtual camera capable of capturing 3000 frames per second. In the third setup, telephoto lenses were used, and each camera had slightly different exposure settings to simulate a camera with a width of 6000 pixels and a high dynamic range. Finally, in the last setup, the cameras were arranged in an arc to capture an outside look from an external perspective. In contrast to their system, our camera system exhibits distinct and unparalleled characteristics. It comprises over 300 cameras, with each camera capturing a 4K resolution. To facilitate efficient background subtraction, we have incorporated a green screen. Our camera system is synchronized through a software protocol, ensuring synchronized operation. While the Stanford Multi-Camera Array captured footage from a limited view angle, our system is capable of capturing the full hemisphere, which is ideal for capturing immersive content for extended reality devices that allow users to freely move and rotate.

Gross *et al.* [12] describe a 3D video acquisition and projection system called Blue-c that uses time multiplexing to switch between acquisition and stereo projection. The surrounding projection screen can be turned opaque for displaying and transparent while recording. The screen shutter is synced with the stereo glasses and the recording cameras. The system uses 16 synchronized cameras and six projectors. Their system is designed for virtual environment collaboration in a large CAVE-like system. In contrast, our system is designed for acquiring content for extended reality devices that require dense sampling of captured volume.

The free viewpoint video system described by Collet *et al.* [5] captures video, computes textured surfaces, and creates free-viewpoint videos for streaming. Their capture system uses 53 high-speed cameras and infrared cameras to capture the visual content of a scene and capture depth information respectively. Their setup has 96 cameras mounted on eight movable stands and 10 mounted overheads to capture the full hemisphere with a capture volume of 2.8 meters in diameter and 2.5 meters in height. The input resolution was 2048 $\times$ 2048, supporting up to 60 frames per second. The system also uses a green screen stage to facilitate efficient background subtraction. Both cameras and color parameters were calibrated to get consistent pixel values across cameras. They use a production lighting kit that can change the lighting system as needed. Their system captures are recorded using PCs where six cameras share a PC. The goal of their project is to create a complete system from capturing to rendering a high-quality video that can be dynamically generated from any view. While their system focuses on mesh generation, our system relies on neural rendering for multi-view rendering. Further, whereas they rely on the captured infrared data to create a mesh, our system generates neural models directly from the input of 300 densely packed cameras.

The capture system employed by Isik *et al.* [13] in the HumanRF project consists of 160 camera arrays. Each camera records at 12 megapixels (4096 $\times$ 3072) and captures 25 frames per second. Their system incorporates a programmable array of 420 LEDs, which are synchronized with the camera shutters. The focus of the HumanRF system is on neural rendering using the camera setup. Their system specializes in capturing close-up views of objects of interest which can be used to capture details around the region of interest. In contrast, our system employs cameras with a significantly larger field of view, with each camera covering objects of interest. Furthermore, our larger capture area can accommodate multiple individuals simultaneously, and can capture complex scenes such as multi-participant performances.

Broxton *et al.* [4] have proposed a system that captures, processes, and streams free-point video for extended reality devices. The capturing system comprises 46 cameras arranged in a hemispherical dome inside looking out. Each camera is frame synchronized and can capture 4K (4096 $\times$ 2160) videos at 30 frames per second. The videos are captured using a very wide field of view to capture and reconstruct occluded areas. Multiple color, alpha, and depth layers are computed from the input and stored in an image atlas. Finally, free-form video is generated to stream for extended reality devices. This system is highly efficient in generating video content for extended reality devices by constraining how much a viewer can move. In contrast, our system captures the full hemisphere, and a viewer can look at content from any direction and position within the capture volume. Unconstrained movement is needed for a collaborative training environment as long as the viewers are within the capture volume.

Schreer *et al.* [31, 32] present a volume capture and streaming system designed for immersive reality devices. The system employs 32 cameras, arranged in 16 pairs of stereo cameras, to capture views from a cylindrical 360-degree perspective. By fusing the collected information, their system generates a single, consistent 3D point cloud data. The volume captured by their system has a diameter of 6 meters and a height of 4 meters. Instead of using a green screen stage, the authors utilize semitransparent diffuse panels with LED lights positioned behind them.

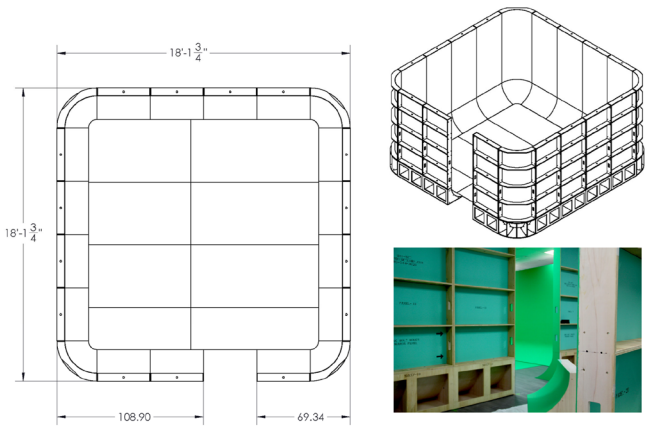


Fig. 2: HoloCamera Cyclorama structure.

The camera employed in their system captures images at a resolution of 20 megapixels and a rate of 30 frames per second. Furthermore, their approach for multi-view 3D reconstruction relies on a vision-based stereo technique.

A previous line of work on camera setups for reconstruction included depth sensors such as Kinect. The Room2Room telepresence system [26] uses three Kinect sensors paired with ceiling-mounted projectors. Maimone and Fuchs [22] fuse five Kinect sensors for room-sized telepresence. Dou *et al.* [7] uses six Kinect cameras grouped into "Panorama Cameras" and "Personal Cameras" to correct for eye gaze. While depth is useful for mesh reconstruction and segmentation, many recent capture studios focus only on RGB cameras due to advancements in RGB cameras outpacing those of depth cameras. For example, while the Azure Kinect DK has a 12 MP color resolution, its depth resolution maxes out at 1 MP in WFOV mode [15]. Under this mode, it only supports a depth range of up to 2.28 meters at 30 Hz. The older Kinect and Kinect V2 used in these systems support an even lower depth resolution of 0.2 MP and have been long discontinued.

2.2 Commercial Studios

There are several volumetric capture studios used for commercial production. They can create content for movies, computer games, extended reality devices, generative chatbots, and other media projects. We examine a few of these studios below and analyze their capabilities.

8i is a professional capture studio that specializes in generating immersive media content, from capturing to streaming [2]. They offer several configurations with up to 60 cameras, each capable of capturing 4K resolution at up to 60 frames per second. They perform global synchronization between cameras to sync shutters and provide several configurable lens packages. The studio is equipped with a green screen stage illuminated by soft lights. They have developed their own calibration software and board for calibration for making captured content consistent across cameras.

4D Views is also a professional recording studio with a capture system that is dedicated to the generation of immersive media content [1, 29]. They offer several configurations, some of which include up to 48 cameras, providing an impressive 200 frames per second, and capturing resolutions as high as 4K. Their capture volume is vast, with a diameter of 6 meters and a height of 3 meters. They have soft lights illuminating their green screen stage.

Metastage [23] is another volumetric capture studio that specializes in creating immersive experiences for movies and extended reality projects. The studio employs 106 cameras, each capable of capturing footage in 4K resolution. A LiDAR system and a motion capture system are incorporated to enhance the capture. Their facilities include a green screen backdrop and are illuminated with LED lights. Their system can capture a volume of 3 meters in diameter and 1.4 meters in height. To create 3D assets, they utilize the capabilities of Microsoft Mixed Reality Capture Studios.

Dimension Studios [6] is a well-known volume capture studio. They employ 110 IOI cameras in conjunction with 140 Nikon Cameras to capture 16K resolution at a frame rate of up to 60 frames per second. The studio boasts a green screen stage equipped with LED lights. To create their 3D assets, they leverage the capabilities of Microsoft Mixed Reality Capture Studios.

Intel Studio is a massive 10,000-square-foot geodesic dome used to capture content for immersive devices [14]. The studio is fitted with 96 cameras capable of capturing at 5K resolution. The studio can do full hemisphere capture and has been used in Hollywood movies.

Sony's volumetric capture technology is another massive capture studio for generating free-viewpoint video [34]. Although they use a large number of cameras, the exact number is not specified.

Our system has 300 densely-packed cameras that capture the full room-sized environment at 4K resolution to provide all the necessary information to challenge and enhance state-of-the-art implicit neural representation models. One problem with existing systems with lower camera density is that neural light field training requires the use of teacher models [36, 40]. Recent work in NeRFs also use teacher models for additional density [8, 27, 28, 39]. With our system's camera density, neural light fields can be trained directly without having to train an entirely separate NeRF teacher model, as demonstrated in prior work [9–11, 18, 19, 33]. Our camera system opens up a path for research towards a future real-time neural light field streaming solution.

3 System Design

In this section, we discuss the construction and configuration of the hardware components of the HoloCamera. We detail specifications of the custom enclosed structure with overhead mounting arrangements, the distributed system of high-performance compute units, the high fidelity camera array sensors, supporting lighting, and the networking and storage infrastructure to interconnect and unify all components of the HoloCamera.

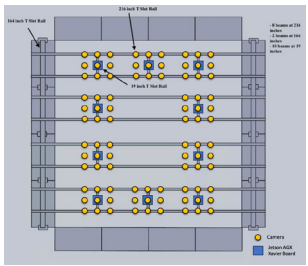
3.1 Physical Structure

Our capture studio infrastructure consists of a custom, free-standing UNISET Cyclorama that provides a seamless green room environment for light-field capture. The custom Cyclorama is made up of 50 individual pieces: 16 UNI-CYC 78' × 40' Flat panels, 16 UNI-CYC 21' × 40' Floor sweeps, four UNI-CYC 90° × 78' Curved panels, four UNI-CYC 21' H × 90° Jewel Inside Sweep corners, and ten UNI-CYC 125' × 46.5' × 94' Cyclorama flooring, all in HDPE dual-sided chroma-key green color. The assembled, sturdy enclosure creates a capture volume of 18' × 18' × 8' enclosed space which: (1) is a free-standing, completely enclosed chroma-key green structure, providing a uniform environment for background subtraction; (2) is modular in design, making the studio portable and flexible for rapid construction and deconstruction; (3) provides the framework for well-kept power distribution, the mounting frame for 16 LED light fixtures, mounting points for 240 individual cameras, mounting locations for 40 Jetson AGX Xavier compute boards, as well as the supporting network infrastructure of Ethernet cables and switches.

To further enhance our volumetric capture capabilities, we designed and constructed a custom overhead camera array to provide a bird's-eye view for the HoloCamera. The rigid body of the Cyclorama provides sufficient structural integrity to support the overhead camera array framework. The overhead camera framework consists of eight 216-inch T-Slot extruded aluminum beams running the full width of the HoloCamera studio, evenly spaced from one end to the other. We securely fasten an additional ten 19-inch T-Slot extruded aluminum beams between the eight 216-inch T-Slot beams to support this framework to provide the mounting points for ten Jetson AGX Xavier compute boards and 60 individual cameras.

3.2 Distributed High Performance Compute Boards

We have deployed 50 NVIDIA Jetson AGX Xavier Developer Kit devices. The AGX Xavier board is comprised of an NVIDIA Volta GPU

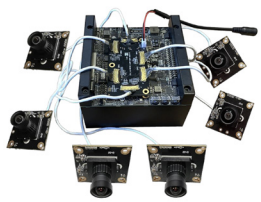


(a) Framework and layout for overhead ceiling cameras

(b) Overhead camera arrangement with angular and directional alignment

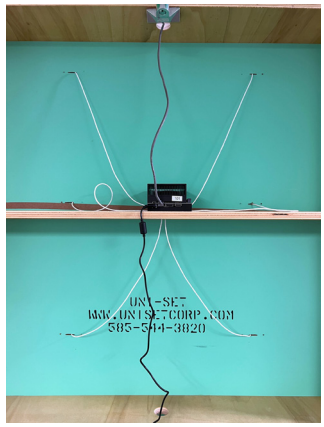
Fig. 3: Overhead camera infrastructure.

(a) NVIDIA Jetson AGX Xavier Developer Kit

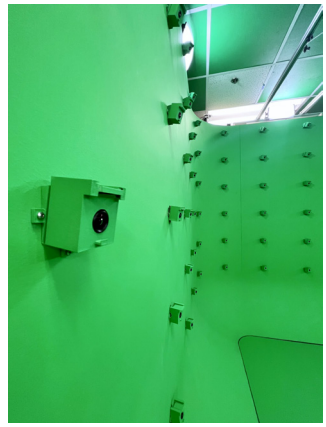


(b) NVIDIA AGX Xavier board with camera adapter board

Fig. 4: NVIDIA AGX Xavier with Leopard Imaging cameras.



(a) Camera module connecting 6 camera sensors to the AGX Xavier board



(b) Camera sensors optimally angled with 3D printed mounts

Fig. 5: Camera sensors and camera modules.

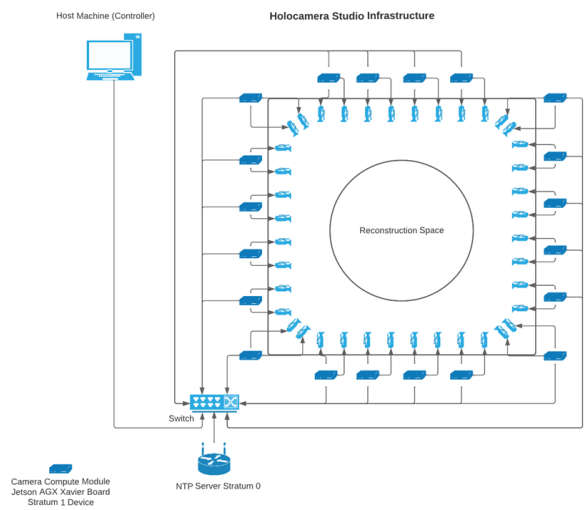


Fig. 6: HoloCamera studio network infrastructure.

with 512 NVIDIA CUDA cores, 8-Core CPU, 32 GB of LPDDR4x memory, and an encoding capability of four 4K streams @ 60 FPS or eight 4K streams @ 30 FPS. To ensure reliability and consistency, in our distributed compute design, we flashed each Jetson AGX Xavier board with a pre-compiled OS image, incorporating the essential libraries, software, dependencies, system, and network configuration settings tailored to support our specific volumetric capture workflow.

3.3 High Fidelity Camera Array Sensors

The full camera sensor array for our studio is comprised of 50 LI-XAVIER-KIT-IMX477M12-H camera modules from Leopard Imaging. Within each module are six 12-megapixel 4K IMX477 sensors with M12 lenses, totaling 300 individual cameras. These modules can capture six video streams at a resolution of 4032×3040 @ 30 fps, 2028×1520 @ 60 fps, and 1920×1080 @ 60 fps, with a field of view (FOV) of 85 degrees (H) × 69 degrees (V) when driven by a single Jetson AGX Xavier board. We have uniformly distributed 240 cameras on the four faces of the Cyclorama structure, with 60 cameras on each face. We have further affixed these cameras to custom 3D-printed mounts, precisely angled toward the center of our capture space to optimize our reconstruction area. The calculation of optimal angles for our cameras is discussed further in subsection 5.4. Each camera is connected via a 500 mm Micro Coax I-PEX Cable, passing through a 25 mm × 4 mm slit in the Cyclorama, and linking to its corresponding LI-JXAV-MIPI-ADPT-6CAM-FP adapter board. We positioned the remaining 60 cameras within the overhead camera array, as described in the previous section, attaching them to omni-directional camera mounts securely fastened to the extruded aluminum rails. We manually adjust each camera to ensure that it is accurately oriented toward the center of the reconstruction volume. Each camera lens is independently focused at the center of the capture volume. Our camera module's IMX477 sensors have an optical format of 1/2.3"; the M12 lenses have 3.9mm focal length and an aperture of F2.8. Given these hardware parameters and focal distances ranging from 9' to 13', the nearest acceptable sharpness distance ranges from 2.33' to 2.53' while the furthest acceptable sharpness is at infinity. Thus all subjects beyond the near focal plane of the cameras are always in focus.

3.4 Networking

Our networking architecture is comprised of separate gigabit Ethernet cables for each Jetson AGX Xavier board, creating a cohesive distributed network of computing modules, and ultimately converging at a local, dedicated 10-gigabit network switch. This switch serves as the central hub for all data transfers within the HoloCamera, facilitating the seamless transmission of hundreds of gigabytes of data over fiber optics to our data center, where it is then processed by a high-performance GPU cluster. The entire network traffic within the HoloCamera is isolated within the space to minimize any outside sources of bandwidth demand.

Local NTP Server

The HoloCamera has a stringent system architecture requirement for precise and highly consistent time synchronization among numerous clients and system processes. To achieve this level of precision, we incorporated a local Network Time Protocol (NTP) Server directly into the 10-gigabit Ethernet switch within the capture studio. This NTP server operates as a stratum 0 source for the Jetson AGX Xavier board clients, elevating them to stratum 1 devices. Notably, this system design intentionally excludes any external time sources, ensuring strict isolation and limiting time information queries exclusively to devices within our system. We discuss software camera synchronization methods we have implemented utilizing NTP protocol in subsection 5.3.

3.5 Storage

To efficiently manage the immense volume of image and video data generated by our volumetric capture system, we took meticulous measures to ensure an ample storage solution. Initially, the image data is stored locally on each individual Jetson AGX Xavier board, with each board equipped with an additional 1TB NVMe M.2 drive directly integrated into the board. This configuration provides a cumulative 50 terabytes of distributed storage right within our capture studio, ensuring rapid access and low-latency retrieval. However, for more streamlined distribution and improved mobility of data to various users, we have implemented a process where the data is subsequently transferred to our 100TB Network Attached Storage (NAS) infrastructure. This NAS setup is strategically provisioned to accommodate the extensive datasets the HoloCamera has the capacity to generate and manage, making it easily accessible to authorized users across our network.

3.6 Lighting

We built the HoloCamera in an academic lab space with general-purpose ceiling lights. While these lights offered consistent overall illumination in the room, they did not provide the uniform light distribution required within our capture studio. To address this, we installed 16 Genaray Spectro LED Essential 500IIB Bi-Color LED Light Panels directly onto the Cyclorama frame. These panels were positioned on top of the Cyclorama, angled toward the center of the reconstruction area to minimize shadows, ensure uniform lighting within the enclosed Cyclorama structure, and allow for adjustable light temperature. Additionally, we placed a light diffuser plate in front of the light panels to control for any harsh or overly bright spots and reduce potential glare.

4 Software System

In this section, we discuss software development and deployment to capture, process, and transfer image data from our cameras. We deployed programs, packages, and scripts to three locations: 1) the Jetson AGX Xavier boards; 2) the central host machine; and 3) the GPU compute cluster.

4.1 Jetson AGX Xavier Board

We use the Libargus Camera API of the Jetson Multimedia API to develop control software for our cameras. The API provides low-level interfaces to acquire images and associated metadata from our cameras. Utilizing this API, we developed our camera capture control executable which enabled us to initialize a capture session, adjust camera parameters, and retrieve captured frames and associated timestamps for synchronization. Adjustable camera parameters include: exposure, analog/digital gain, white balance, capture duration, and capture mode (image/video). We coordinate exposure and analog gain adjustment to capture moving subjects with reduced motion blur while maintaining consistent exposure. Auto white balance lock and manual white balance gain adjustments allow us to maintain consistent white balance across all cameras for non-RAW image captures. We use the timestamps associated with captured frames to synchronize the encoded frames across cameras on each Jetson AGX Xavier board; the method is discussed more in section 5.3. We acquire RAW images by initializing the output with STREAM_TYPE_EGL stream type. EGL is a Khronos interface that provides efficient transfer of images between APIs. Using this stream type, we are able to set the capture pixel format which supports RAW 16-bit image encoding. FFmpeg [35] runs on all Jetson AGX Xavier boards to perform onboard transcode operations including RAW image Debayering (subsection 5.2), image profile adjustment, and image format conversion.

4.2 Central Host Machine

A Linux workstation acts as the central host machine to control all the cameras. We developed a custom GUI for easy access to all main camera functions as shown in Figure 7. The GUI provides a user-friendly interface to facilitate and streamline dataset captures. In the underlying framework, we have a series of scripts that use SSH protocol to call into functions located on each of the Jetson AGX Xavier boards. The core functions include: initializing NTP synchronization, configuring remote data storage locations, dispatching parallel capture signals, converting data formats, and retrieving captured data. Benefiting from our distributed computing infrastructure, all processes are automatically parallelized. Additional utility functions are also implemented to restart Jetson AGX Xavier boards, facilitate checking camera synchronization, and verify data acquisition completeness.

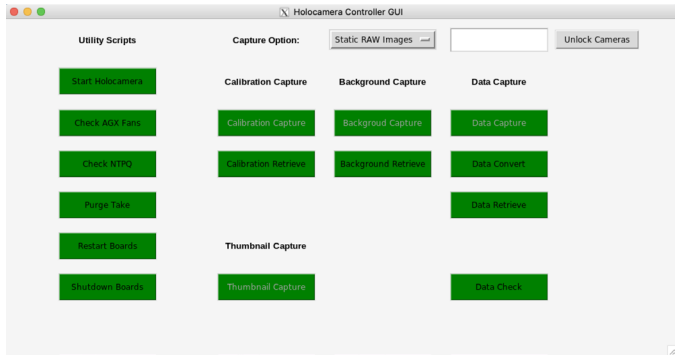


Fig. 7: HoloCamera controller GUI: intuitive interface to control camera captures as well as associated utility functions.

4.3 Compute Cluster

We have built a GPU compute cluster consisting of eight Nvidia RTX 2080 Ti GPUs to further preprocess our captured datasets for training and reconstruction. A modified and distributed version of BackgroundMattingV2 [20] is deployed on the cluster to quickly extract the foreground in all of our datasets. BackgroundMattingV2 performs robust background matting using a two-segmentation network architecture, benefiting from the additional GPU computing resources available in the cluster. The uniform green background has minimal background visual features to ensure that a high-quality background matting is generated with each capture. The background matting can be used to optimize NeRF [24, 25] model training as well as more conventional 3D reconstruction methods such as visual hull [16]. To preview the capture for each dataset, we have integrated Instant-NGP [25] to generate a NeRF and render a thumbnail alongside the generated background matting. Since both background matting and model training occur on the same computer cluster, we optimize the performance by adapting both steps to avoid unnecessary memory copies and disk access. Our deployment of Instant-NGP reuses RGB views loaded into memory and generated background matting saved in memory.

5 Camera System

In this section, we focus on efforts related to configuring our cameras. We describe designs and choices made to calibrate, position, and synchronize our camera array.

5.1 Camera Calibration

We perform camera calibration for all cameras by setting up the capture volume with feature-rich calibration patterns. We utilize the Structure-from-Motion (SfM) pipeline from COLMAP to calibrate the extrinsic and intrinsic parameters of our cameras [30]. We use the radial camera model during calibration which accounts for focal distance, principal point, and radial distortions.

Collet *et al.* [5] described a calibration Octolith to facilitate their calibration process. To provide a rigid calibration target, we constructed a custom calibration monolith shown in Figure 9 consisting of a 20×20×34-inch box stacked on top of a 24×24×34-inch box. We then mounted calibration patterns to cover all surfaces of the structure. The Octolith calibration required captures of a few frames at different positions. Our calibration process requires only a single static capture of our calibration setup.

During the early stages of our experiments, we printed 8.5” × 11” letter-size checkerboard patterns and mounted them similar to the placement on the Octolith. Our testing configuration had four prints on

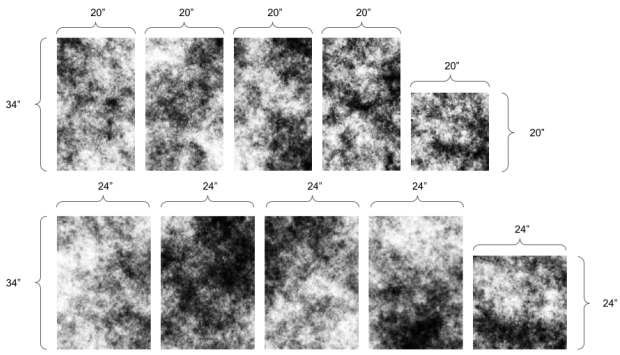


Fig. 8: Distinct random noise patterns covering all visible surfaces of the calibration monolith; every calibration surface has its own unique random noise pattern; this optimized calibration pattern design drastically increased feature count and track length in camera calibration.

Fig. 9: Calibration monolith object with optimized random noise patterns, captured by one of the cameras from HoloCamera.

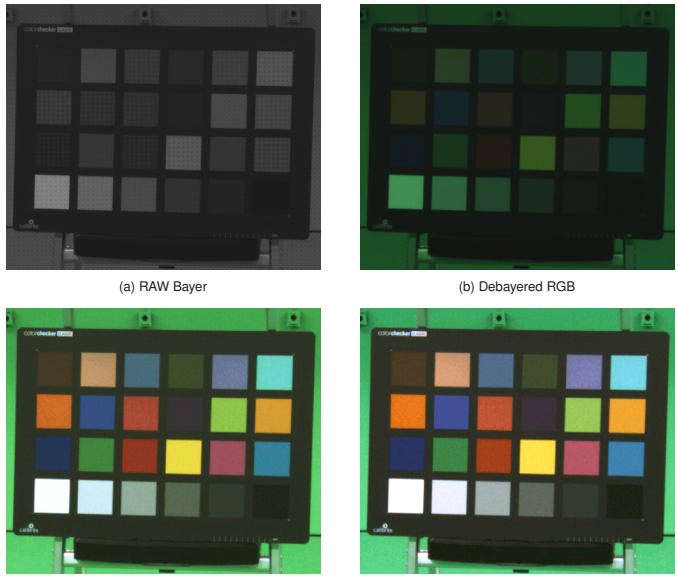


Fig. 10: Color calibration from RAW sensor output: (a) RAW Bayer RGGB export from image sensor (b) RAW image Debayered into RGB image (c) Manual adjustment (d) Color calibration using ColorChecker.

each side surface of our calibration monolith. To further refine our calibration quality, we found that checkerboard patterns are designed for single or stereo camera calibration in which the entirety of the calibration pattern must be visible and recognized from camera views. For multi-camera calibration using the SfM pipeline, particularly with a volumetric capture studio, only a subset of the calibration targets are seen by a camera due to difference in viewpoints and occlusion. Thus, the main factor affecting reconstruction accuracy is having a generous amount of clearly-defined features on multiple surfaces. This increases the number of feature observations from each camera view and results in a longer track length for each of the 3D points reconstructed in the scene. Li *et al.* [17] presented a feature descriptor-based calibration pattern specifically designed for SIFT [21] descriptors. They experimented with calibrating stereo cameras and a four-camera system which resulted in more features and less reprojection error compared to chessboard calibration patterns. We followed the same principle but scaled the number of calibrated cameras to 300. We used the associated toolbox published in the OpenCV [3] library to generate distinct random noise calibration patterns. We tested our theory by replacing the letter-size checkerboard patterns with newly generated random noise patterns printed on the same letter-size paper. The placement of the test random patterns are identical to test checkerboard patterns. We recorded calibration statistics for this setup in Table 1 as Test Random Pattern. We observed marginal improvements in the number of observations and mean reprojection error. To further optimize the random noise pattern for maximized calibration accuracy, we then printed full-size flat surface posters with uniquely generated random noise patterns covering all visible surfaces of our calibration monolith (Figure 9). A single random noise pattern large enough to cover the calibration monolith was generated, then it was sectioned into the printed posters shown in (Figure 8). This guarantees all calibration surface patterns are unique. The resulting calibration had comparably less mean reprojection error with a drastic increase in the number of observations and mean track length. Calibration statistics from above-mentioned calibration pattern designs are shown in Table 1.

5.2 Color Calibration

To achieve the highest image quality for our static captures, we export RAW images from the cameras in Bayer RGGB format. This ensures we have raw sensor data without any loss of quality through any default encoding process. We perform raw image processing and color calibration to achieve the highest quality data obtainable through our system. The raw images contain luminance values received by the photosensor through a Bayer filter. We Debayer the raw image using FFmpeg [35] pixel format conversion, and then perform a series of image adjustments to the brightness, contrast, exposure, and saturation.

Finally, we perform color correction using a Macbeth ColorChecker board as reference colors. The ColorChecker is widely used in photography, television, and printing for color profiling. It consists of 24 squares of standardized color patches with very tight colorimetric tolerances. Images of the board captured under different environments serve as a series of reference colors. We place the ColorChecker in the center of the capture volume and capture a static photo using our raw image capture pipeline. OpenCV's color correction model recognizes the ColorChecker board and calculates correction values to adjust white balance and color responses relative to the reference colors. We then incorporate the color correction values into our capture pipeline and apply them to all captures to restore images to their correct color.

5.3 Camera Synchronization

Camera synchronization is essential for capturing frame-synced multi-camera visual data. Out-of-sync frames introduce ambiguity and error to reconstruction algorithms using our dataset. For our system setup, we implement a two-tier synchronization scheme to synchronize cameras on each Jetson AGX Xavier board and cameras across different Jetson AGX Xavier boards respectively. Through a combination of hardware and software synchronization, we are able to achieve a tight bound of centisecond synchronization accuracy.

Table 1: Calibration quality comparison: Test checkerboard patterns and test random noise patterns are printed on letter-sized paper. Each side surface of the calibration monolith had four of the same letter-sized pattern (checkerboard or random noise pattern) mounted. Optimized random noise patterns are unique full-size prints mounted on each side of the calibration monolith to maximize calibration accuracy.

	Test Checkerboard Pattern	Test Random Pattern	Optimized Random Patterns
Points	80,769	**97,141**	91,212
Observations	567,612	675,212	**1,134,301**
Mean track length	7.03	6.95	**12.44**
Mean observations per image	1892	2251	**3781**
Mean reprojection error	0.82	**0.74**	0.76

Onboard Synchronization

Each of our 50 Jetson AGX Xavier boards drives six cameras through a Leopard Imaging adapter board. The adapter board provides hardware shutter synchronization between the six cameras through a primary and secondary configuration. One camera, set up as the primary camera, triggers the other five camera shutters. This hardware synchronization scheme ensures all cameras connected to the same Jetson AGX Xavier board are capturing frames with microsecond accuracy. However, we found that despite synchronized camera shutters, the encoded frames could still be out of sync. Upon close examination of the associated timestamps, we discovered the multithreaded encoding process introduced a constant offset to the encoded frames' timestamps because threads are initialized sequentially. To correct this systematic error, we employed a software scheme mimicking the hardware primary-secondary configuration to ensure encoded frames on each Jetson AGX Xavier board are in sync. There are six consumer threads initiated on a Jetson AGX Xavier board to encode frames from each of the six cameras. Each thread will request frames from its camera in the order that the frame is captured. Each frame has a timestamp recording the time it was captured. We set the primary camera as reference and use the timestamp from its frame as a benchmark to synchronize the other five secondary cameras. If the timestamp from the frame of a secondary camera is ahead of the benchmark timestamp, we preserve that frame and place it back in the frames queue so it will be retrieved in the next iteration. If the timestamp is behind the benchmark timestamp, we discard the current frame and retrieve frames until we find the timestamp that is in sync with the primary camera's frame. Because hardware synchronization ensures a synchronized camera shutter, we are guaranteed to have all cameras retrieving frames with the same timestamp. We run this synchronization algorithm at the beginning of every capture session until all cameras converge. We observed that the cameras converge within a few frames of the cameras initializing. Thus, frames captured by cameras attached to the same Jetson AGX Xavier board will capture and encode frames with timestamps synchronized with microsecond accuracy.

Time Synchronization Across Jetson Boards

With all six cameras connected to each Jetson AGX Xavier board synchronized down to microseconds, the problem of synchronization is reduced to synchronizing all 50 Jetson AGX Xavier boards. To synchronize the boards, we leveraged the connected network structure among the boards and utilized Network Time Protocol (NTP) to synchronize the boards with sub-millisecond accuracy. We established an additional Raspberry Pi board to serve as a local NTP server. The Raspberry Pi board is connected directly to the network switches that connect the Jetson AGX Xavier boards. All 50 Jetson AGX Xavier boards establish NTP synchronization on stratum 1 with the NTP server located at stratum 0. Having such a close network setup enforces minimum network connection delay. We measured the NTP statuses of the Jetson AGX Xavier boards to have less than 0.5 milliseconds round trip delay time, less than 0.1 milliseconds offset to the server's clock, and a jitter of less than 0.1 milliseconds. These values validate that we have sub-millisecond synchronization between the internal clocks of all the Jetson AGX Xavier boards on our network system. We then synchronize all camera captures by implementing time gate logic on top of the manufacturer-provided Application Programming Interface (API). For each capture session, we specify that all the cameras use their connected Jetson AGX Xavier board's system time as reference, and only start capture once a time gate has been reached. With this programming logic, we guarantee dispatching camera capture signals across all 300 cameras with sub-millisecond precision.

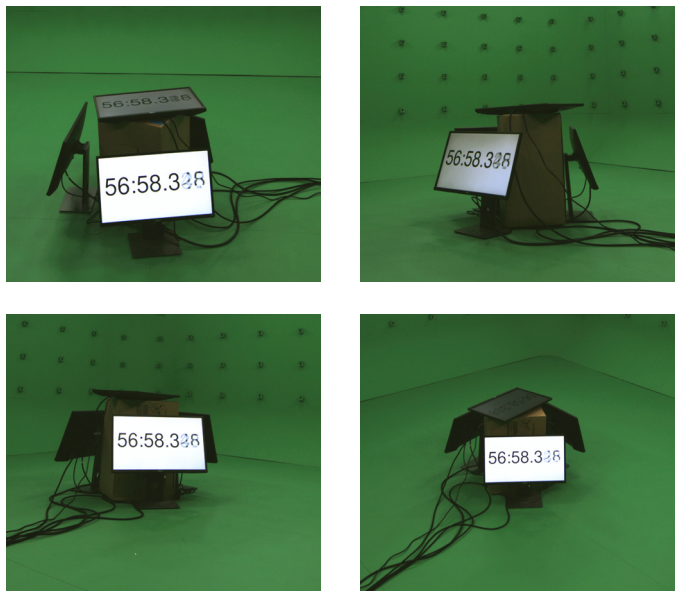
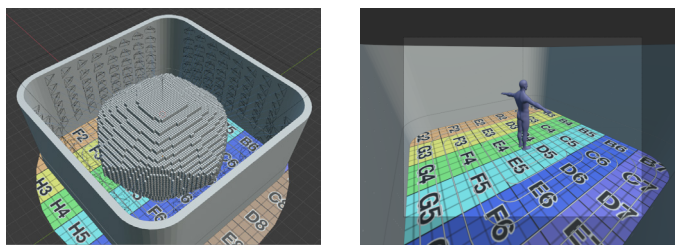


Fig. 11: Four monitor setup displaying stopwatch at 60Hz, Capture showing blended digits.



(a) Digital twin software with intersecting volume via voxel-based approach

(b) Digital twin software with view perspective from virtual camera

Fig. 12: HoloCamera digital twin software.

Synchronization Accuracy Measurement

We had implemented a two-tier hardware and software synchronization scheme to enforce camera start capture signals to be sent to all cameras within one millisecond. However, there are lower-level processes in the camera control API that interface directly with the hardware layer. Once the start capture signal has been sent to the camera boards, there is a delay before the cameras start acquiring frames. The variation in delay between different camera boards could introduce offsets, which

Fig. 13: Gallery of a subset of the released dataset: Screenshot of rendered frames from trained NeRF models.

lead to synchronization errors. To further quantify the synchronization accuracy of our camera setup, we devised a novel method to find the tightest error bound of camera synchronization.

Our setup consists of four computer monitors displaying a stopwatch timer. All four monitors are connected to a single workstation so their contents are mirrored and in sync from each other. These monitors are arranged so that any camera can observe the stopwatch on at least one monitor. The timer has millisecond precision, but the monitors have a refresh rate of 60 Hz, resulting in the stopwatch timer values having 16.67 ms intervals between timestamps, i.e. if the number shown currently is x seconds, the next number will be $x+0.01667$ seconds. We set the cameras to capture still images with an exposure time of 16.67 ms, matching the refresh interval of the monitor. However, because the cameras are not synchronized precisely with the monitor's refresh rate, the resulting images often exhibit a blend between two consecutive stopwatch numbers displayed on the monitors. As a consequence, captured timestamps appear blurred after the 0.1-second digit. Nevertheless, based on examination of the blended imagery, we have confirmed that all cameras are capturing the same two consecutive stopwatch numbers. Therefore, we can conclude our cameras are at most off by one frame between two consecutive refreshed numbers. Consequently, the measured tightest error bound in our camera synchronization measurements is 16.67 ms. Our actual synchronization error is in theory closer to the millisecond precision we have achieved across all boards. With a higher refresh rate timestamp source, we will be able to validate higher precision synchronization within our system of cameras.

5.4 Camera Angle Optimization

An important task in the development of the 3D volumetric video capture studio was computing the optimal angles of the 240 video camera modules affixed to the perimeter walls of the studio. We used the 3D modeling software, Blender, to build a digital twin of the studio. Using Blender's camera modeling framework, we simulated each camera's view frustum, factoring in their specific properties such as sensor dimensions, resolution, focal length, and lens distortion. By projecting the simulated camera frustums into the digital studio model, various configurations of cameras were evaluated for the intersecting volume of all projected camera frustums, with the goal of maximizing the effective overlapping capture volume of the studio.

To determine a measure of intersecting volume, a voxel-based approach similar to visual hull was used. The digital studio space was partitioned into fixed-size voxels. Each camera's view frustum was projected into the voxel volume. Any voxel outside of the intersection of all camera frustums was discarded as outside of the capture volume. The total volume of interior voxels was measured for each configuration to determine the optimal angular combination of camera modules. We used the optimal values to model camera mounts tailored for each of the 240 individual cameras. We then 3D printed the custom-designed camera mounts affixed each camera to its dedicated mount.

6 Lightfield Dataset Release

To foster future research in volumetric capture, we are releasing a total of 30 cinematic-fidelity light-field datasets. We show the images from a subset of our release in Figure 13. Each dataset consists of static lossless images from all 300 cameras, their respective foreground masks, and calibrated camera extrinsic and intrinsic parameters. The dataset covers a variety of individuals and groups of human subjects. We also include human interaction with pets, instruments, candles, and sports equipment. HoloCamera's high-resolution acquisition pipeline results in a variety of distinctive visual features in our dataset. These include elements such as soft dress fabric, view-dependent reflective surfaces (such as Mylar balloons), candle flame, fine geometry elements (such as violin strings and tennis racket mesh), and dog fur covering a wide range of scenarios that often pose challenges for multi-view reconstruction and rendering.

Foreground masks have been processed to facilitate reconstruction efforts. We also export our camera calibration results from COLMAP to include associated camera rotation and position, focal length, principal point, and distortion parameters. These camera parameters can be easily incorporated into any reconstruction pipeline.

The dataset is hosted at `https://holocamera.umd.edu`.

7 Conclusion

We stand at the threshold of a new era in building virtual environments of cinematic quality, offering immersive experiences from any chosen perspective. Although the current generation of virtual environments can effectively convey photorealistic details for a broad range of objects and scenes, their portrayal of humans does not attain the same level of quality or realism, and often fails to elicit a suspension of disbelief. This paper presents the end-to-end pipeline for building a room-sized volumetric capture system with 300 cameras. This paper will enable others to rapidly build similar volumetric capture studios and tailor them to their specific applications. While light fields offer an information-rich medium for static and dynamic scenes, a significant barrier to their widespread adoption is a lack of sufficiently rich, finely calibrated dataset to enable research into fast and compact representations of such high-dimensional data. Our public release of a structured light field dataset should catalyze meaningful research into new algorithms for efficient storage, editing, and streaming of such datasets.

ACKNOWLEDGMENTS

We sincerely thank the anonymous reviewers for their valuable suggestions to improve the paper. This work has been supported in part by the NSF Grants 18-23321 and 21-37229, and the State of Maryland's MPower initiative. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the research sponsors.

REFERENCES

[1] 4Dviews. 4Dviews - Volumetric Video Capture Technology. https: //www.4dviews.com/volumetric-systems, Sept 2023. 3

[2] 8i. ASPX Hologram Stages. https://8i.com/aspx-hologram-stages/, Sept 2023. 3

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 6

[4] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Trans. Graph.*, 39(4), aug 2020. doi: 10.1145/3386569.3392485 2

[5] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 2, 5

[6] Dimension. Studio Overview. https://www.dimensionstudio.co/studios, Sept 2023. 3

[7] M. Dou, Y. Shi, J.-M. Frahm, H. Fuchs, B. Mauchly, and M. Marathe. Room-sized informal telepresence system. In *2012 IEEE Virtual Reality Workshops (VRW)*, pp. 15–18, 2012. doi: 10.1109/VR.2012.6180869 3

[8] S. Fang, W. Xu, H. Wang, Y. Yang, Y. Wang, and S. Zhou. One is all: Bridging the gap between neural radiance fields architectures with progressive volume distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 597–605, 2023. 3

[9] B. Y. Feng and A. Varshney. SIGNET: Efficient neural representation for light fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14224–14233, 2021. 3

[10] B. Y. Feng and A. Varshney. Neural subspaces for light fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1685–1695, 2024. doi: 10.1109/TVCG.2022.3224674 3

[11] B. Y. Feng, Y. Zhang, D. Tang, R. Du, and A. Varshney. PRIF: Primary ray-based implicit function. In *European Conference on Computer Vision*, pp. 138–155. Springer, 2022. 3

[12] M. Gross, S. Würmlin, M. Naef, E. Lamboray, C. Spagno, A. Kunz, E. Koller-Meier, T. Svoboda, L. Van Gool, S. Lang, K. Strehlke, A. Vande Moere, and O. Staadt. Blue-c: A spatially immersive display and 3d video portal for telepresence. *ACM Trans. Graph.*, 22(3):819–827, jul 2003. doi: 10.1145/882262.882350 2

[13] M. Işık, M. Rünz, M. Georgopoulos, T. Khakhulin, J. Starck, L. Agapito, and M. Nießner. HumanRF: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. doi: 10.1145/3592415 2

[14] Intel. Intel Studios Showcases Volumetric Production at 77th Venice International Film Festival. https://www.intel.com/content/www/us/en/newsroom/news/studios-volumetric-production-venice-film-festival.html, Sept 2023. 3

[15] G. Kurillo, E. Hemingway, M.-L. Cheng, and L. Cheng. Evaluating the accuracy of the Azure Kinect and Kinect v2. *Sensors*, 22(7), 2022. doi: 10.3390/s22072469 3

[16] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994. doi: 10.1109/34.273735 5

[17] B. Li, L. Heng, K. Koser, and M. Pollefeys. A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1301–1307, 2013. doi: 10.1109/IROS.2013.6696517 6

[18] D. Li, B. Y. Feng, and A. Varshney. Continuous levels of detail for light field networks. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 3

[19] D. Li and A. Varshney. Progressive multi-scale light field networks. In *2022 International Conference on 3D Vision (3DV)*, pp. 231–241, 2022. doi: 10.1109/3DV57658.2022.00035 3

[20] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8762–8771, 2021. 5

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, nov 2004. 6

[22] A. Maimone and H. Fuchs. Real-time volumetric 3d capture of room-sized scenes for telepresence. In *2012 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4, 2012. doi: 10.1109/3DTV.2012.6365430 3

[23] Metastage. Metastage : Our Tech. https://metastage.com/our-tech, Sept 2023. 3

[24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[25] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. doi: 10.1145/3528223.3530127 5

[26] T. Pejsa, J. Kantor, H. Benko, E. Ofek, and A. Wilson. Room2room: Enabling life-size telepresence in a projected augmented reality environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, p. 1716–1725. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2818048.2819965 3

[27] C. Reiser, S. Peng, Y. Liao, and A. Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14335–14345, 2021. 3

[28] C. Reiser, R. Szeliski, D. Verbin, P. Srinivasan, B. Mildenhall, A. Geiger, J. Barron, and P. Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 3

[29] B. Rogez. SIGGRAPH 2023: 4Dviews unveils cutting edge volumetric capture system HOLOSYS+. https://3dvf.com/en/siggraph-2023-4dviews-unveils-cutting-edge-volumetric-capture-system-holosys/, Sept 2023. 3

[30] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4104–4113, 2016. 5

[31] O. Schreer, I. Feldmann, P. Kauff, P. Eisert, D. Tatzelt, C. Hellge, K. Müller, S. Bliedung, and T. Ebner. Lessons learned during one year of commercial volumetric video production. *SMPTE Motion Imaging Journal*, 129(9):31–37, 2020. 2

[32] O. Schreer, I. Feldmann, S. Renault, M. Zepp, M. Worchel, P. Eisert, and P. Kauff. Capture and 3d video processing of volumetric video. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4310–4314, 2019. doi: 10.1109/ICIP.2019.8803576 2

[33] V. Sitzmann, S. Rezchikov, B. Freeman, J. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021. 3

[34] Sony. Volumetric Capture Technology That Goes Beyond Omnidirectional Visualization. https://www.sony.com/en/SonyInfo/technology/stories/Volumetric_Capture/, Sept 2023. 3

[35] S. Tomar. Converting video formats with FFmpeg. *Linux Journal*, 2006(146):10, 2006. 5, 6

[36] H. Wang, J. Ren, Z. Huang, K. Olszewski, M. Chai, Y. Fu, and S. Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, pp. 612–629. Springer, 2022. 3

[37] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics (TOG)*, vol. 24, pp. 765–776, 2005. 2

[38] B. S. Wilburn, M. Smulski, H.-H. K. Lee, and M. A. Horowitz. Light field video camera. In *Media Processors 2002*, vol. 4674, pp. 29–36. SPIE, 2001. 2

[39] Y. Wu, X. Li, J. Wang, X. Han, S. Cui, and Y. Lu. Efficient view synthesis with neural radiance distribution field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18506–18515, 2023. 3

[40] H. Yu, J. Julin, Z. A. Milacski, K. Niinuma, and L. A. Jeni. Dylin: Making light field networks dynamic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12397–12406, 2023. 3