VR Social Copresence with Light Field Displays

NATHAN MATSUDA, BRIAN WHEELWRIGHT, JOEL HEGLAND, and DOUGLAS LANMAN, Facebook Reality Labs Research, USA

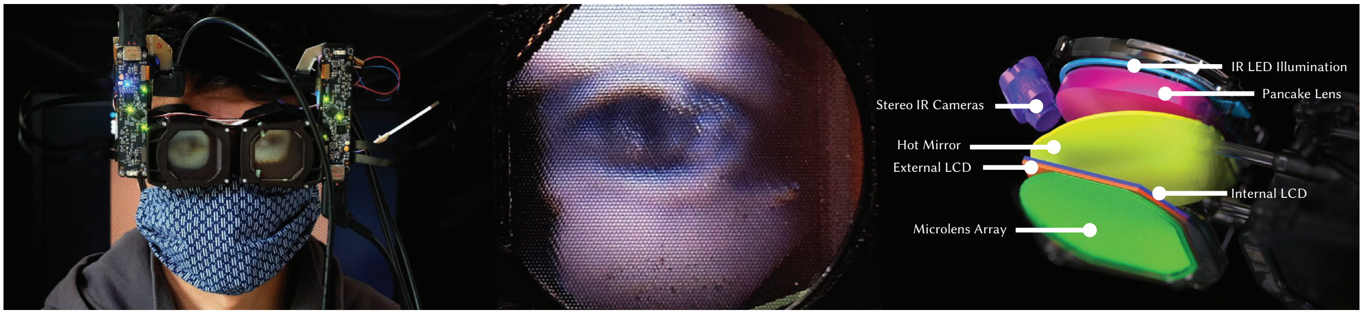


Fig. 1. *Left:* Our light field reverse pass-through VR prototype depicts a live, three-dimensional reconstruction of a user's eyes that is visible to an arbitrary number of external viewers. *Middle:* These reconstructions are depicted using an autostereoscopic light field display with sufficient resolution to accurately depict the user's gaze. *Right:* The headset contains two display modules, each comprising: a ring of infrared LEDs (light blue) providing illumination for facial reconstruction without interfering with the visible-wavelength VR display, a pancake lens viewing optic (pink) enabling a compact form factor, a dichroic hot mirror (yellow) folding the optical path for a pair of infrared eye-tracking cameras (purple), an inward-facing VR display LCD (blue), a world-facing LCD (orange), and a microlens array (green).

As virtual reality (VR) devices become increasingly commonplace, asymmetric interactions between people with and without headsets are becoming more frequent. Existing video pass-through VR headsets solve one side of these asymmetric interactions by showing the user a live reconstruction of the outside world. This paper further advocates for *reverse pass-through VR*, wherein a three-dimensional view of the user's face and eyes is presented to any number of outside viewers in a perspective-correct manner using a light field display. Tying together research in social telepresence and copresence, autostereoscopic displays, and facial capture, reverse pass-through VR enables natural eye contact and other important non-verbal cues in a wider range of interaction scenarios, providing a path to potentially increase the utility and social acceptability of VR headsets in shared and public spaces.

CCS Concepts: • **Computing methodologies** → **Virtual reality**.

Additional Key Words and Phrases: virtual reality, light field displays, social copresence, reverse pass-through

ACM Reference Format:
Nathan Matsuda, Brian Wheelwright, Joel Hegland, and Douglas Lanman. 2021. VR Social Copresence with Light Field Displays. *ACM Trans. Graph.* 40, 6, Article 1 (December 2021), 13 pages. https://doi.org/10.1145/3478513.3480481

1 INTRODUCTION

To date, research into virtual reality (VR) displays has largely focused on reaching parity with direct-view displays, including improving resolution, reducing latency, mitigating vergence-accommodation conflict, and developing more comfortable form factors. Yet, as emphasized by Gugenheimer et al. [2019], little attention has been paid to resolving another core deficiency: VR displays isolate the user from their environment and, in doing so, limit VR use and acceptance in shared and public spaces [Mai and Khamis 2018; Schwind et al. 2018]. Eliminating this isolation is a key motivation for the development of *video pass-through VR*, wherein the VR headset user sees a reproduction of their external environment and the individuals within it. Yet, a crucial gap remains: External viewers cannot hold a natural conversation with a VR headset user, whose upper face and eyes remain occluded.

Several efforts have been made to depict the occluded features of a VR headset user's face on external, world-facing displays. Chan and Minamizawa [2017] depict a stylization of the user's eyes—driven via eye tracking—to give a sense of the user's gaze direction and attention. Their approach stops short of depicting the surrounding facial regions and eliminates all perspective depth cues. To partially address these limitations, Mai et al. [2017] depict a hand-tuned face model that is aligned to the perspective of a single external viewer and supports a manually controlled gaze direction. Yet, across these and related works, we identify remaining capabilities that are necessary to deliver authentic social copresence using external, world-facing displays, including faithful reproduction of the entire occluded region of the face, accurate depiction of three-dimensional depth, and support for multiple external viewers.

We recently introduced *reverse pass-through VR* [2021] to meet these needs. In this paper we further detail the motivation, implementation, and evaluation of this concept. Akin to how pass-through VR delivers realistic stereoscopic imagery to a single headset user, we advocate for user-facing cameras, real-time facial reconstruction, and autostereoscopic world-facing displays to deliver an accurate

Authors' address: Nathan Matsuda; Brian Wheelwright; Joel Hegland; Douglas Lanman, Facebook Reality Labs Research, USA.

0730-0301/2021/12-ART1
https://doi.org/10.1145/3478513.3480481

recreation of the user's hidden face and eyes for an arbitrary number of external viewers (see Figure 1). While the underlying technologies have been under development for decades, we are aware of no effort to combine them in this manner to unlock social VR. Furthermore, our proposed system is timely, leveraging recent research and industry trends in VR facial capture [Lombardi et al. 2018], high-resolution autostereoscopic displays [Martínez-Corral and Javidi 2018], and more compact VR headsets [Maimone and Wang 2020] (which further improve resolution with world-facing autostereoscopic displays). If successfully developed, reverse pass-through VR may increase the social acceptability of VR to better compete with augmented reality (AR), which already allows direct eye contact via *optical see-through* displays. In this manner, users may benefit—in public—from the wider fields of view and better occlusion cues currently delivered with VR displays.

1.1 Contributions

Our primary technical contributions include the following:

- We motivate the use of external light field displays to achieve authentic social copresence between a VR headset user and multiple external viewers. We review prior work on VR facial capture and autostereoscopic displays, then link a geometric model predicting apparent gaze errors, a user study, and the need to support multiple external viewers to establish the need for light field-based reverse pass-through VR.
- We provide extended implementation details for a proof-of-concept reverse pass-through VR prototype using currently available camera and display components. We also present a software facial reconstruction and light field rendering pipeline capable of running in real time on a desktop PC.
- We evaluate the performance of this prototype in terms of geometric gaze error and show qualitative results that demonstrate real-time, three-dimensional facial capture and display. We further evaluate perceived gaze accuracy with a preliminary remote user study.

1.2 Overview of Benefits and Limitations

Reverse pass-through VR is closely related to prior work on autostereoscopic displays for telepresence applications. As shown by Nguyen and Canny [2005] and Pan and Steed [2014], such systems more accurately depict gaze and attention for multiple viewers than flat panel displays. Furthermore, external viewers perceive accurate motion parallax and binocular depth cues, which have been shown by Kim et al. [2012] to significantly increase the sense of social presence. Beyond these user studies establishing the benefits of autostereoscopic displays for group telepresence, we emphasize that a reverse pass-through headset may ultimately appear no different than an ordinary pair of eyeglasses: The user's eyes simply appear to be framed by any remaining opaque portions of the headset.

These benefits do not come without limitations. Foremost, autostereoscopic displays often exhibit limited spatial resolution. As assessed in Section 3.2, our light field display trades between spatial and angular resolution. While other autostereoscopic displays exist, this trade-off is common to the thin three-dimensional display architectures that are compatible with increasingly compact VR headsets. Due to this limitation, the resulting images of the face and eyes appear at significantly reduced resolution than would be possible by directly viewing the underlying LCD (see Figure 1). As noted by Lombardi et al. [2018], fine facial features, such as eyelashes, pores, and vellus hair, all contribute to the perception of human emotions. Thus, future reverse pass-through VR systems will benefit from adopting higher-resolution autostereoscopic displays.

Beyond the limitations imposed by reduced spatial resolution, our reverse pass-through VR prototype is currently hindered by our selected facial capture technology. As described in Section 4.2, we implement a machine-learning-based stereo reconstruction framework—optimized per user—achieving a limited field of view and frame rate (lower than our light field display), exhibiting depth reconstruction errors, and requiring colorizing images captured by infrared cameras. Furthermore, our system does not support relighting of the face by environmental sources, which Palmer et al. [2020] have shown may be important for estimating gaze direction. Taken together, future reverse pass-through VR systems stand to benefit most by adopting enhanced VR facial capture technologies.

2 RELATED WORK

Reverse pass-through VR is closely related to prior work on telepresence displays, which aim to establish a sense of shared space bridging dissimilar environments. Within this field, perceived gaze direction in images of the face has long been studied, including for painting and illustration [Wollaston 1824] and for television [Anstis et al. 1969]. Over the last century, human-computer interaction and computer graphics researchers have studied perceived gaze using interactive digital communication systems. This work has established the central challenges for the field, which include achieving direct eye contact (and correct gaze direction in general), accurately reproducing the three-dimensional appearance of participants, and faithfully conveying attention and emotional state. In this section we review prior telepresence work spanning system engineering, psychophysics, and human-computer interaction design.

Facial Projection Mapping. The concept of displaying interactive graphics on the face of a real performer was explored in the HyperMask project [Yotsukura et al. 2002]. While we aim to authentically reconnect the user's outwardly visible identity with their hidden face, HyperMask decouples facial identify from the performer. Variations on this idea include telepresence via facial projection onto mannequins [Lincoln et al. 2009b; Schubert et al. 2012] as well as modifying facial appearance directly [Bermano et al. 2017]. Prior facial projection work also delves into the the impact of gaze geometry on social interactions [Al Moubayed et al. 2012; Moubayed et al. 2012], which is a primary motivation for our work.

Telepresence Gaze Cues. The depiction and perception of gaze direction has been studied for a wide variety of display technologies, including actuated 2D displays [Sirkin et al. 2011] and spherical displays [Pan and Steed 2014]. In the latter work, Pan and Steed introduce an autostereoscopic gaze-preserving telepresence system and later test, via user studies, the accuracy of perceived eye contact [Pan and Steed 2016]. Their findings, particularly relevant here, establish that perspective-correct imagery significantly improves perceived eye contact over a 2D image, with binocular depth cues further improving the viewer's perception of gaze direction.

Autostereoscopic Display of Humans. The topic of autostereoscopic displays—also known as "glasses-free 3D TVs"—covers a wide range of techniques and applications, including a significant body of work devoted to systems reproducing human gaze [Jones et al. 2009; Kim et al. 2012], as well as human bodies and faces more generally [Gotsch et al. 2018; Lincoln et al. 2009a; Nagano et al. 2013].

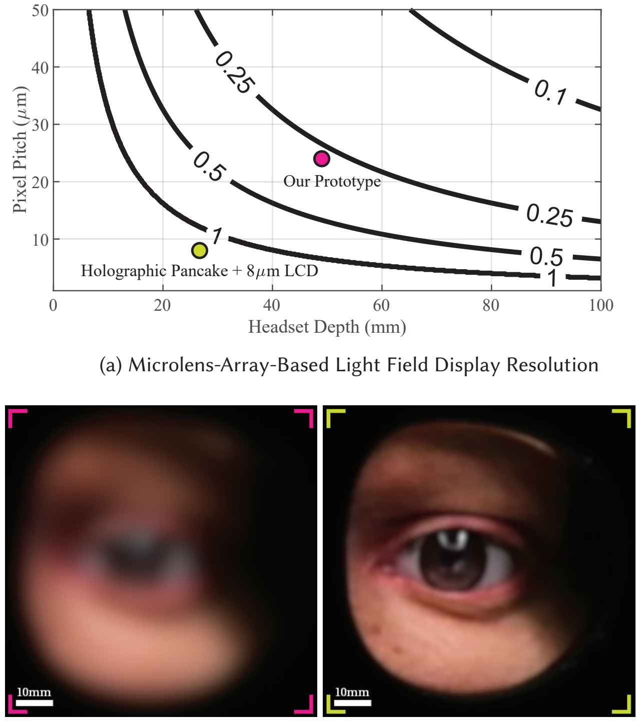
Autostereoscopic Head-Mounted Displays. Autostereoscopic displays have also been demonstrated in the context of head-mounted displays, including Lanman and Luebke [2013] and Huang et al. [2015].

VR Face Replacement. Facial capture for telepresence has been recently explored in the context of VR. Such efforts typically employ face tracking within a head-mounted display (HMD) to produce realistic facial geometry [Li et al. 2015; Olszewski et al. 2016], and then map prerecorded images of the VR headset user's face over an external camera view of the user [Frueh et al. 2017; Thies et al. 2018; Wei et al. 2019]. Alternatively, facial reconstructions can be produced entirely using deep learning, such as with an autoencoder architecture [Lombardi et al. 2018], or generative adversarial networks [Wang et al. 2019b]. Orts-Escolano et al. [2016] briefly describe a similar facial replacement technique in augmented reality telepresence, including the importance of gaze cues for collaborative scenarios.

World-Facing Head-Mounted Displays. In closely related work, Misawa and Rekimoto [2015] use a head-mounted tablet—worn by a surrogate—to support embodied telepresence for a remote third party. We emphasize that the majority of such prior work focuses on telepresence, while we are concerned with copresence. As previously discussed, Chan and Minamizawa [2017] use an eye tracker to drive a pair of cartoon eyes, presented on a mobile phone mounted to a VR headset. While not explicitly aiming for correct eye gaze, Gugenheimer et al. [2019] and Wang et al. [2020] both provide context about the headset user's virtual environment via externally mounted screens and projectors. We note that such displays are complementary to our system, whereas AR glasses do not present a similar ability to use a world-facing display for multiple modalities. Tying back to the start of this section, Furukawa et al. [2019] use projected images, together with an actuated mannequin, to visually and physically couple the experience of external participants to that of the VR user. Mai et al. [2017] explore the scenario we are interested in—mitigating an asymmetric interaction between an individual wearing a headset, surrounded by individuals not wearing headsets. While they did not find significant benefits in perceived attentiveness and comprehension when displaying facial images, it should be noted that their implementation used a 2D mobile phone display showing a fixed model.

3 REVERSE PASS-THROUGH LIGHT FIELD DISPLAYS

Prior work in social telepresence establishes the need for autostereoscopic displays to accurately reproduce a user's eye gaze and facial appearance from the vantage point of every observer. Reprojection to a 2D display, as was previously demonstrated by Mai et al. [2017], is insufficient due to incorrect binocular and motion parallax depth cues. These prior findings lead us to set a design requirement that reverse pass-through VR systems must use an autostereoscopic display, otherwise external viewers will need to don special eyewear (e.g., tracked shutter glasses or VR headsets), further hindering the practicality and social acceptability of VR displays in shared and public spaces. Many autostereoscopic displays have been proposed, but, as we establish in this section, we advocate that microlens-array-based light field displays are most compatible with the current research and industry trends toward more compact form factors.



(a) Microlens-Array-Based Light Field Display Resolution

(b) Our Prototype (Modeled)

(c) Holographic Pancake (Modeled)

Fig. 2. As described in Section 3.2, the apparent resolution of a reverse pass-through light field display depends on three factors: the resolution of the underlying 2D display, the optical properties of the microlens array (MLA), and the distance from the light field display to the user's face and eyes. (a) The display pixel pitch and headset thickness are varied as the MLA remains fixed. Contours denote the effective light field resolution (in cycles/mm) at the depth coinciding with the user's face and eyes. (b) We model the apparent resolution using the prototype described in Section 4.1 and (c) using the holographic pancake architecture proposed by Maimone and Wang [2020] together with a recent high-density microdisplay [Kopin 2021]. Thus, we anticipate rapid improvements are possible by following recent research and industry trends.

3.1 Research and Industry Trends in VR Displays

Recent advances in polarization-based optical folding or "pancake"' viewing optics establish a path toward significantly reduced headset volumes [Geng et al. 2018; Wong et al. 2017]. Recently, Maimone and

Wang [2020] have shown that holographic optics may further decrease headset thickness, approaching sunglasses-like form factors. Maintaining such thin enclosures, while supporting the addition of reverse pass-through VR displays, means that large-form-factor autostereoscopic displays, such as volumetric or multi-projector displays, do not present compelling design choices. Furthermore, the world-facing display should also be sufficiently bright to be visible in varying lighting conditions, while consuming low enough power to be head-worn as a mobile device. These restrictions further rule out light-attenuating approaches such as multilayer displays. We do not rule out the possibility of using a holographic display in the future, but no existing digital holographic display yet approaches the form factor, étendue, and framerate possible with microlens-array-based light field displays.

We observe that light field displays have not been widely adopted, in part, due to their inherent spatio-angular resolution trade-off, which effectively limits the "depth of field" (i.e., the range over which a high-resolution image can be depicted). However, given current state-of-the-art display panel resolutions (with densities are being driven by the VR industry itself), emerging VR viewing optics, and the limited depth of field needed to support an image plane at a headset-user's eye relief, light field displays are well suited for constructing compelling reverse pass-through VR systems today.

3.2 Light Field Display Resolution and Depth of Field

A light field display multiplexes four spatio-angular dimensions into the two dimensions supported by flat panel displays. The relationship between effective spatial resolution and virtual image distance is determined by the display pixel pitch, the microlens array (MLA) pitch, and the MLA focal length [Lanman et al. 2013]. Two of these are fixed by VR industry trends: the pixel pitch and the virtual image distance (i.e., the displacement from the front of the headset to the user's nominal eye position). We note that the MLA pitch, focal length, field of view, and elemental image size are interrelated, and also depend on the refractive index of the microlens material.

In Figure 2, we assess how the progression toward thinner headsets and higher-density displays will impact the resolution of reverse pass-through VR. Here we have fixed the microlens parameters to the values outlined in Section 4. This analysis predicts a spatial resolution approaching 0.25 cycles/mm, which we confirm via measurements in Section 4. Figure 2 further includes visualizations of the predicted resolution, aligning with our experimental results. We also predict the resolution for future reverse pass-through light field displays using holographic pancake optics [Maimone and Wang 2020] and a high-density display (pitch under 10μm) that should be available in the near future [Kopin 2021]. This combination should support a spatial resolution around 1 cycle/mm, showing that significant near-term gains may be achieved, further establishing reverse pass-through VR as a practical, visually compelling solution that is within reach.

4 IMPLEMENTATION

Our physical prototype design uses components that are practically plausible and, in most instances, already available off the shelf. We designed hardware subsystems (the light field display and stereo camera systems), and software subsystems (eye image synthesis, camera and display calibration, and light field rendering) following this principle. We hope that this motivates others to pursue research on reverse pass-through topics in the near term.

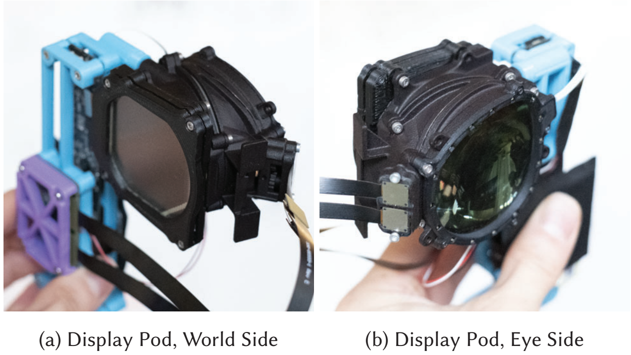

(a) Display Pod, World Side (b) Display Pod, Eye Side

Fig. 3. (a) External view of prototype display pod, showing LCD-MLA stack. The LCD driver is mounted in the blue chassis at left, with a short flex cable connecting the LCD. Stereo cameras are connected via flex cables exiting the pod at right of photo and connecting to the stereo camera driver board. (b) Rear view showing the IR LED ring and pancake lens. Stereo cameras are more clearly visible at the left of this view.

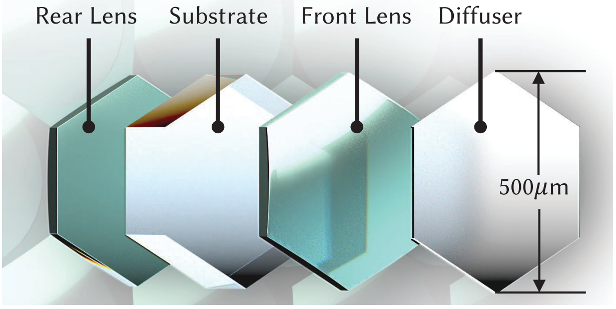
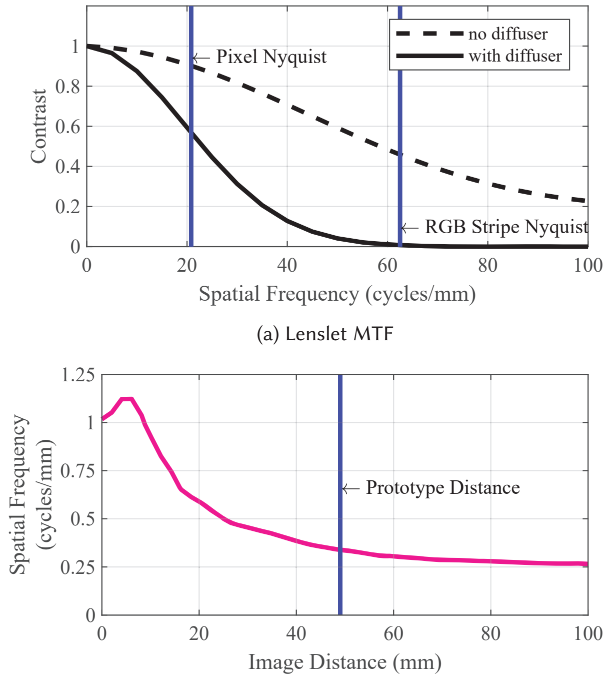


Fig. 4. Lenslet stack consists of two mirror-image surfaces molded in resin on a 200μm Corning EagleXG substrate, resulting in an approximately 330μm vertex-to-vertex thickness. The resulting lens has an effective focal length of 520μm and an f-number of F/1.04 for the edge-to-edge limiting aperture. A 2° FWHM engineered diffuser imposes a cutoff between Nyquist frequencies for the LCD's 24μm RGB pixel grid and 8μm RGB stripe pattern to prevent the microlens from resolving individual RGB subpixels.

With the trend toward thinner VR headsets in mind, we designed a reverse pass-through module around an overall track length typical for pancake lenses. Each module contains a VR display and pancake viewing optic, an external light field display, and an eye capture subsystem (See Figure 3). The headset contains two such modules, one for each eye, rigidly mounted to each other via carbon fiber rods over the nose bridge and affixed to the head with the head strap from an Oculus Go headset.

4.1 Light Field Display

We established in Section 3 that MLA-based light field displays are uniquely well suited to this application. We back up this argument



(a) Lenslet MTF

(b) Measured Light Field Depth Of Field

Fig. 5. (a) A Zemax simulation of the lenslet modulation transfer function for the 0° tangential field. The bare lenslet MTF is denoted by a dashed line, while the solid line denotes the diffused MTF. Nyquist frequencies for the RGB superpixel and underlying RGB stripe patterns are shown as labeled vertical lines. Adding a 2° FWHM engineered diffuser maintains contrast about 50% for RGB superpixel resolution while cutting contrast to near zero for individual RGB subpixels. (b) Light field Display resolution as a function of image distance for our experimental prototype measured by capturing sinusoids with periods ranging from the display width down to 2 pixels horizontally. Cutoff taken at 50% Michelson contrast. This measurement agrees with the prediction in Section 3.

by implementing a high performance MLA design using an off-the-shelf LCD and a commercially mature resin-molding manufacturing technique. We then show that this display can be driven at real-time rates with a conventional game rendering engine.

To maximize field of view for a given focal length, the MLA lenslets need to be as fast as possible (high numerical aperture) while maintaining imaging performance. A dual-side MLA distributes the optical power over two surfaces, reducing aberrations compared to single-side MLAs. To simplify fabrication and reduce cost, the two surfaces are designed with the same optical prescription. The resulting lens design, shown in Figure 4, uses a hex-packed lenslet pattern with each lenslet supporting a 42° field of view, an effective focal length of 520μm, and a 500μm pitch between lenslet centers. The lenslets are F/0.90 when measured corner-to-corner, and F1.04 when measured edge-to-edge.[1] Holographix LLC lithographically produced a master and replicated the dual-sided MLA on a 200μm glass substrate (Corning EagleXG). The MLA polymer is Holographix proprietary with n_D=1.693 ; v_D = 29.8.

The Zemax-predicted modulation transfer function for a single lenslet is shown in Figure 5a for the 0° tangential field. Labeled vertical lines in the plot show the spatial frequencies associated with the display's RGB superpixel Nyquist frequency and that of the underlying RGB stripe. The lenslet MTF (dashed line) retains near 50% contrast for the 0° field at the stripe frequency, which will produce distracting sharp images of the RGB subpixels rather than the aggregate color of all three. We selected a 2° diffuser from Luminit to introduce a low-pass filtering on the stripe spatial frequencies. The solid line in the plot denotes the MTF with diffuser installed. With the diffuser, contrast remains above 50% for the 0° field at the superpixel spatial frequency, but drops near zero at the stripe frequency. In addition to defocusing the subpixels, the diffuser breaks up Moiré effects caused by beating between the subpixel pitch and the MLA pitch. We opted to install the diffuser as a separate layer in the MLA stack (see Figure 4) in order to compare performance with and without it, but the diffuser structure could be incorporated directly into the MLA surface.

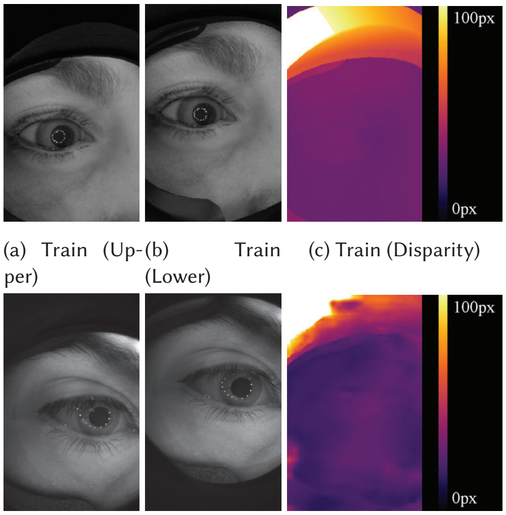
The microlens array is backed by a BOE 1600×1600 color LCD with 24μm pixel pitch (8μm RGB stripe). We drive this display with a Synaptics VXR7200-based display bridge. A duplicate of this display bridge drives the second, inward-facing LCD, which is the same 1600×1600 panel, placed directly back-to-back with the outward-facing LCD.

To confirm the depth of field we predicted in Section 3 (Figure 2), we measured the spatial frequency cutoff of our fabricated light field displays by sweeping horizontal sinusoids with periods ranging from the full display width down to 2 pixels at virtual images planes ranging from the physical display plane to 100mm behind the display. We then found the spatial frequency threshold where Michelson contrast drops to 50%. The resulting frequency cutoff as a function of target image distance is shown in Figure 5b. The measured resolution at the eye plane matches our prediction, around a third of a cycle per millimeter at the headset's nominal eye relief.

4.2 Eye Image Capture and Synthesis

Continuing with our tactic of using existing techniques where possible, we designed an eye capture system that is compatible with existing eye tracking architectures. We use a pair of near-IR Omnivision OV9281 CMOS sensors driven by an Omnivision OV580 USB stereo capture board. A Chroma Technology T700 IR-reflective hot mirror provides a more on-axis view (17.5deg off-axis) than would be possible with an around-the-lens camera configuration. See Figure 1 for a cutaway view of the camera geometry.

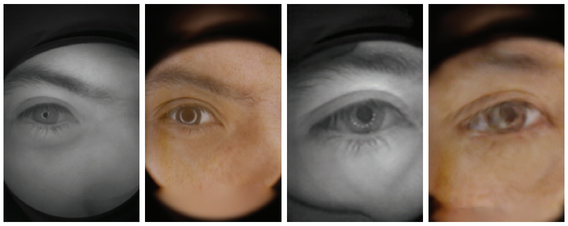
[1] We also produced a version of the microlens array with embedded black chrome aperture mask on the glass substrate. This aperture mask serves two functions: 1) mitigate cross-talk modes between neighboring lenslets and 2) mitigate stray light from imperfect cusps between neighboring lenslets. The apertures are 400μm diameter over a matching 500μm hex-pack (F/1.45). However, we found that dark regions surrounding the eye images (due to the geometry of the eye-cups and stereo camera field-of-view in our physical prototype) served to limit cross-talk and stray light in the non-masked MLAs. We chose to use the non-masked, full aperture version of the MLA, which produced brighter images, for the results in Section 5.



(a) Train (Upper) (b) Train (Lower) (c) Train (Disparity)

(d) Test (Upper) (e) Test (Lower) (f) Test (Disparity)

Fig. 6. Upper (a) and lower (b) synthetic renders, with geometry matching the vertically-oriented stereo eye cameras, are shown along with ground truth synthetic disparity (c). These images are used to train AnyNet for depth inference. At run-time, the live stereo upper (d) and lower (e) view of the eyes are passed to the network, which returns an inferred disparity map (f). The cameras are pre-calibrated to enable stereo rectification prior to inference, and to support reprojecting the recovered disparity to mesh geometry.



(a) Train (IR) (b) Train (Color) (c) Test (IR) (d) Test (Color)

Fig. 7. Captured training pairs, with registered IR image (a) and RGB color image (b), are used to train a CycleGAN network. Live IR views of the face (c) are passed to the network at run-time, which returns an inferred RGB color image (d).

The stereo pair are calibrated with a conventional OpenCV pipeline, using a small printed circle grid pattern.

The requirement for high quality, high framerate, 3D reconstruction and colorization from infrared images for a limited domain (eyes) leads naturally to deep learning techniques. We surveyed leading candidates for stereo depth inference and colorization and selected AnyNet [Wang et al. 2019a] and CycleGan [Zhu et al. 2017], respectively.

We produced a training dataset with a face model derived from the Digital Emily project [Alexander et al. 2009] implemented in Blender [Blender Foundation 2021]. The model was re-textured using captured color and IR images of the author's face, then rigged to randomly span a range of positions relative to the headset, eye gaze directions, and eyelid poses. The resulting dataset contains 10,000 examples of left and right images for color and IR textures.

This dataset was used to refine the AnyNet model, pre-trained on the SceneFlow Driving Dataset [Mayer et al. 2016], over 300 epochs using default parameters. Example inputs and depth inferences for synthetic and captured images are shown in Figure 6.

To train the colorization network, we captured 300 real IR/color pairs by affixing a color camera (Arducam IMX298) to the headset eye cup such that the entrance pupil was co-located with the left IR camera entrance pupil. The CycleGan model was trained on these images for 200 epochs. Example IR inputs and color inferences are shown in Figure 7. Due to the low resolution of the light field image at the eye plane (described in Section 3.2), colorization is performed on images downsampled to 160×96, which exceeds the maximum spatial resolution possible for this display (see Section 3.2).

These networks are of course overfit to one target user. In a research setting, this can be expanded to multiple users, but we would expect that a future reverse pass-through system at scale could generalize using neural avatars in the manner of Lombardi et al. [2018].

4.3 Calibration and Runtime Operation

The manufacturing accuracy of the LCD and MLAs described in this section greatly reduces the number of parameters that need to be calibrated for each display in practice. The lenslet prescription, pitch, and tiling are known precisely. The LCD pitch and tiling are also known precisely. Three degrees of freedom can be fully accounted for on the software side: positioning in the plane of the display and clocking. This leaves two out-of-plane rotational degrees of freedom and separation between LCD and MLA that must be manually accounted for during assembly. We used 0.003" shim stock from Artus, placed in at 3 positions around the perimeter of the display stack, then affixed the MLA to the display with epoxy resin. During this process we displayed a white image on the LCD. Since the diffuser is not yet installed, correct MLA focus is sharply distinguished when the RGB subpixels come into focus. While not exact, we find this in practice to be repeatable thanks to the consistent flatness of the LCD cover glass and MLA substrate.

Using the real-time renderer described in Appendix A, we sweep horizontal and vertical offsets until a displayed cross-hair is level and centered in the display. There may be some performance gains to be had by performing active alignment during display assembly rather than manual placement and tuning. Doing so would require extensive physical tooling and automation, so we leave this to others interested in production engineering.

We also performed a rudimentary color calibration of the display by stepping through grey levels 0-255 and capturing them with the same exposure settings used for all measurements in Section 5, then used the inverse of the resulting curves as a color lookup after the color inference step.

During live inference, camera images are captured from the OV580 board via a Python wrapper. Stereo rectification is performed with OpenCV using the pre-calculated camera calibration. The rectified images are passed to AnyNet and CycleGAN, running as

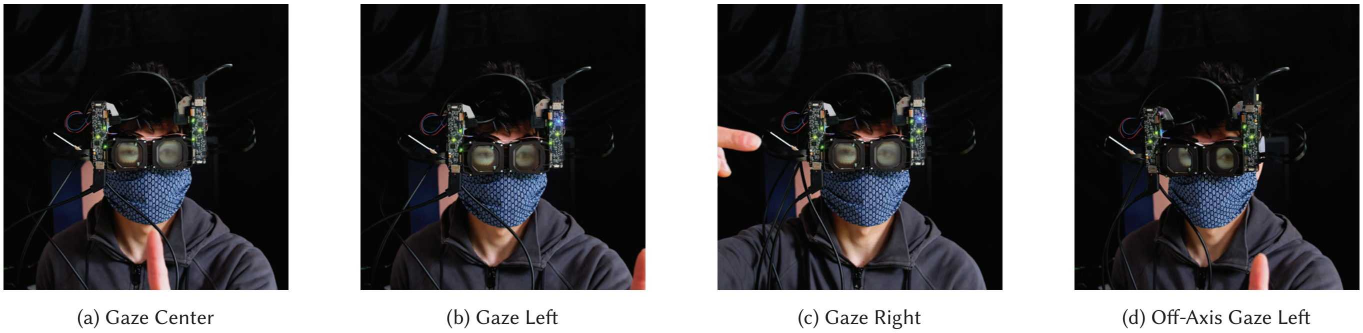


Fig. 8. Photos of a subject using the full headset with both eyes in operation. To demonstrate perspective-correct gaze reproduction, we show the subject looking at their finger in the center (a), to the left (b), to the right (c), and with the head turned off-axis, gazing to the left (d). View our supplementary video to see this sequence in motion.

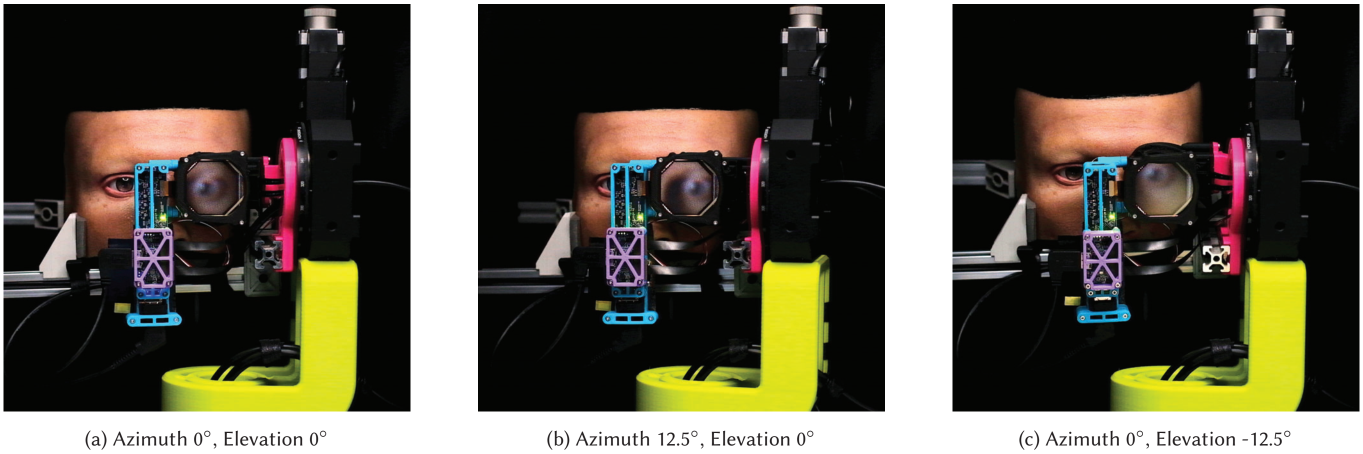

(a) Azimuth 0°, Elevation 0°
(b) Azimuth 12.5°, Elevation 0°
(c) Azimuth 0°, Elevation -12.5°

Fig. 9. Our experimental setup with display pod and driver boards mounted to a 2-axis gimbal, along with a mannequin head used a static facial reference. The centered position is shown in (a), rotated 12.5° to camera left in (b), and rotated 12.5° down in (d). The light field display is showing the offline (photogrammetry) reconstruction of the mannequin in these images.

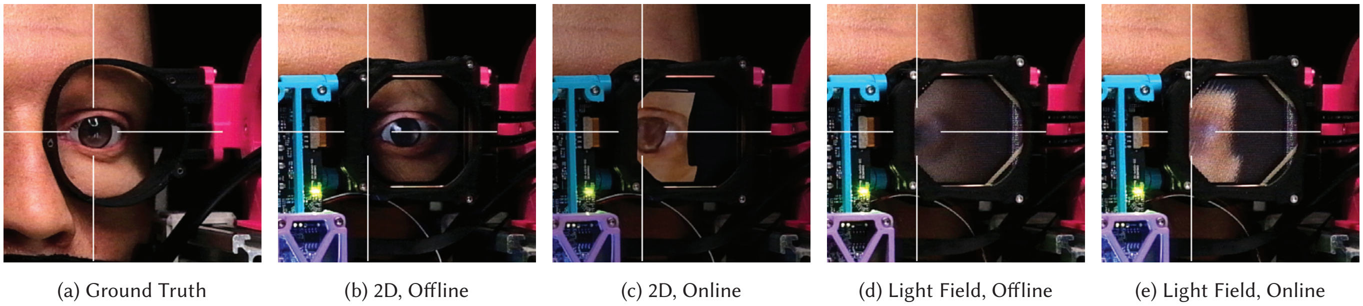


Fig. 10. (a) Ground truth view of the mannequin head with display assembly removed. (b) Baseline 2D approach with offline photogrammetry reconstruction of face. (c) Baseline 2D approach displaying live view using example view synthesis approach outlined in Section 4.2. (d) Proposed light field architecture displaying offline photogrammetry reconstruction. (e) Proposed architecture displaying live view using example view synthesis approach outlined in Section 4.2. White cross-hairs on all images are aligned to the ground truth pupil position for reference. All images were captured with the head at -12.5° azimuth.

separate processes, using ZeroMQ [Hintjens 2013]. The inferred disparity output from AnyNet is projected to 3D points using the previous camera calibration. Both processes then pass the images to texture buffers in Blender, also using ZeroMQ (adapted from Heindl et al. [2020]). In Blender, the built-in RGB-to-XYZ modifier is used to displace the points of a subdivided plane to their inferred physical locations, and that plane is texture mapped with the inferred color. The screen-space refraction renderer described in Appendix A then produces the images for display using Blender's built-in EEVEE renderer.

This prototype runs at interactive rates as measured on an an Intel i7-8700k Desktop with dual Nvidia Titan Xp graphics cards. Anynet and CycleGAN were restricted to one of the GPUs, while Blender/EEVEE ran on the other, which was attached via DisplayPort to the LCD screen at its native resolution (1600×1600).

We measure the average framerate of each thread individually: AnyNet - 17fps, CycleGAN - 19fps, Blender - 46fps. Note that there is no inter-process synchronization; if a new XYZ map or color map is not available, the renderer presents the previously received buffers. It goes without saying that these performance numbers could be significantly improved with careful optimization, but we believe the high performance we attained with research-quality Python scripts further supports the practicality of our system.

5 RESULTS

Images of the wearable headset running the live inference reconstruction are shown in Figure 8. A subject gazes toward a finger held center, left, and right, as well as to the left with an off-axis head position, to show perspective-correct reproduction of eye images.

Of course, results depicting an autostereoscopic headset are poorly represented on a 2D page or even a 2D video, but we recommend viewing our supplementary video as the perspective-correct parallax enabled by the light field display is more evident when viewed in motion.

5.1 Comparison with 2D Projection

We noted in Section 3 that light field displays sacrifice resolution to produce autostereoscopic images. Does the added depth accuracy outweigh the lost resolution when comparing to tracked 2D displays? Using a tracked 2D display will not work for most situations where there is more than one person in the room, but a light field alternative should nonetheless provide social cues that are as good or better in the single-viewer case. A key way that tracked 2D displays fail in the single-user case is the lack of stereoscopy; the reprojected view can be correct for one eye, or the average position of the eyes, but not both eyes simultaneously.

We implemented the 2D tracking approach for comparison with our prototype hardware by swapping in the same LCD, but without the MLA. We mounted one full display pod to a two-axis gimbal comprised of two Thorlabs HDR50/M rotation stages and a number of 3D printed support assemblies. We also mounted a silicone mannequin head fabricated by Legacy Effects to use as a static, and thus repeatable, reference face. Using the known gimbal angle and camera location, we render the correct perspective for a view 32mm to the right of the camera, as if the camera were the left eye on a person with 64mm interpupillary distance. We later compare to the light field output for that same angle. To control for reprojection errors due to the online facial reconstruction pipeline, we also produced an offline photogrammetric model of the mannequin.

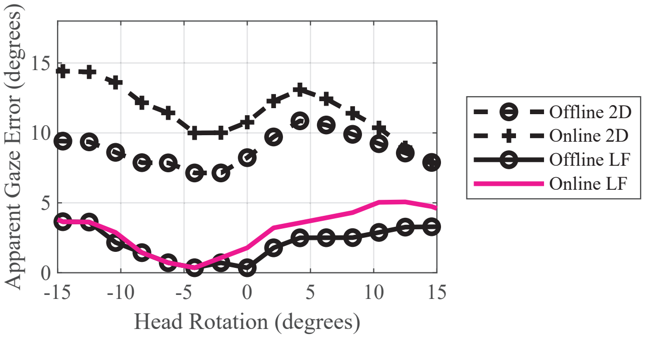


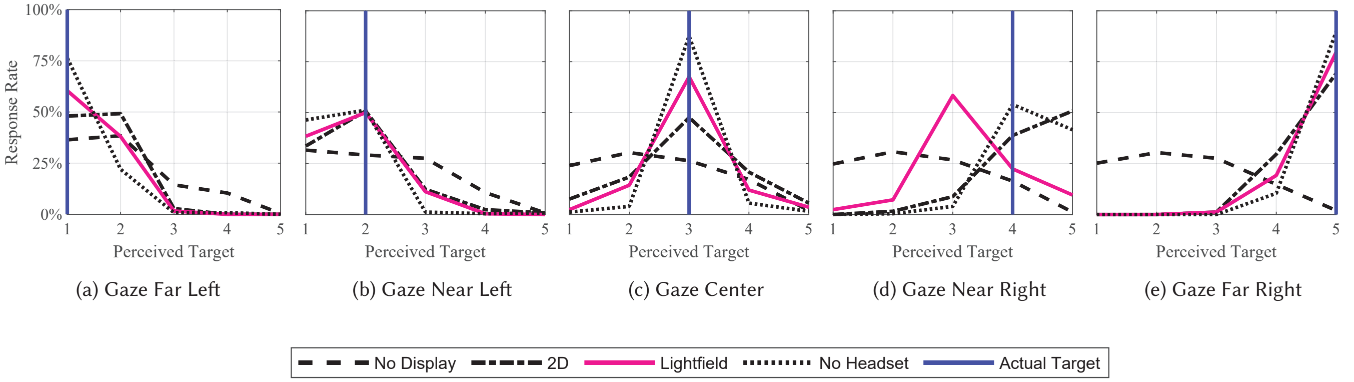
Fig. 11. Pupil displacement in degrees (See Appendix B for details), plotted as a function of mannequin head azimuth for offline and online modes (o-ticks and +-ticks, respectively). 2D reprojection (dashed lines) consistently shows angular errors that are much higher than the light field display (solid lines). While the 2D reprojection asymmetry can be partially accounted for by the reprojection targeting one eye to the right of the observing camera, there is also asymmetry present in the light field modes, which can best be explained by miscalibration of the eye capture camera extrinsics with respect to the headset and light field display.

Figure 10 depicts three of these display modes, along with a ground truth image of the mannequin without the eye-pod mounted, all from a viewing angle of 12.5° azimuthal rotation from the axis of the display. While the offline 2D display shows significant displacement of the pupil due to the parallax induced by the LCD's approximately 50mm offset from the plane of the pupil, the offline and online light field displays show correct perspective.[2]

Chen [2002] finds that people perceive eye contact when the pupil is displaced by no more than 1° up, left, and right of the view axis, and no more than 5° down. Using the user study results from that paper as a guide, we confirm that our proposed architecture produces autostereoscopic eye images with pupil displacements that are smaller in magnitude than the equivalent 2D projection method, despite the reduction in resolution due to the light field display. To do this, we approximate the apparent gaze error with a pupil reprojection model detailed in Appendix B. In short, the model approximates gaze direction as the vector originating at the eyeball center and pointing toward the projection of the pupil image onto the eyeball using the tracked pupil position[3], known eye relief (15mm), headset thickness (35mm), eyeball radius (12mm), and the known head angle. Using this model entirely sidesteps perceptual

[2] Results were captured using a Fujifilm X-T2 camera with 35mm lens set to F2.2, ISO 1250, with an exposure time of 1/30s. The camera was placed 1m from the display plane of the prototype (a plausible conversational distance), with the optical axis centered on the middle of the display.

[3] Tracking the relatively low resolution, low contrast images of the light field display is challenging, so we captured additional image sequences where the input color was replaced with a white circle over black background corresponding to the mannequin iris. We then tracked this blob in the captured images and use the centroid as an estimate for the pupil position.



(a) Gaze Far Left

(b) Gaze Near Left

(c) Gaze Center

(d) Gaze Near Right

(e) Gaze Far Right

Fig. 12. Histogram plots showing participant-reported perceived gaze target for a user wearing a blank external display (dashed line), live reconstruction on a 2D display (dash-dot line), live reconstruction on our proposed light field display (red solid line), and wearing no headset (dotted line). Gaze targets lie along a line centered 1m in front of the headset-wearer, with the following targets: a) 500mm left of camera, b), 250mm left of camera, c) the camera itself, d), 250mm right of camera, and e), 500mm right of camera. The light field display outperforms the 2D and blank display cases in the extreme right and left, and most importantly in the center, which emulates eye contact. Gaze detection is ambiguous in the intermediate 250mm positions, even for the no headset case, indicating the limitations of this remote user study approach (see Section 5.2 for details.)

Table 1. One-way Analysis Of Variance (ANOVA) results for each of the five gaze targets (left column) show that when comparing the 2D reprojection mode to the no-headset case (middle column), response distributions have statistical significance, suggesting users could tell the difference, for 3 of the 5 targets. On the other hand, when comparing the light field displays to the no-headset case (right column), responses for only one target showed statistical significance, suggesting users could not distinguish between the light field and no-headset gazes for the other 4 targets. Statistically significant entries, where the *p*-value is less than 0.05, are bolded.

Target	*p*-value, 2D & None	*p*-value, LF & None
1	**0.0000**	0.0551
2	**0.0004**	0.1220
3	0.9514	0.9888
4	0.9929	**0.0000**
5	**0.0012**	0.2170

effects in exchange for a simple geometric model that allows repeat gaze error approximations to be measured on our motorized gimbal.

Figure 11 shows errors for 2D reprojection and light field display modes, using both the offline photogrammetry reconstruction of the mannequin head and live pass-through. The light field display output produces more accurate view perspectives than the tracked 2D output, meeting the 1° criteria for some head rotation angles. Extrapolating from the improved light field resolution predicted for future headsets in Section 3.2, we expect light field reverse pass-through displays to consistently provide reprojection errors below this threshold in the future. Not only do the non-light field pupil rotation errors exceed this threshold everywhere, those approaches only work for a single person.

5.2 Preliminary User Study

Institutional COVID policies prevented us from doing in-person evaluation of our prototype at any point in the past 15 months, so we had to fall back to a remote user study. We aimed to establish that perceived eye gaze would be more accurate for the light field display compared to the equivalent 2D display or a baseline headset with no external display. The test conditions were inspired by Pan and Steed [2014]: the headset wearer gazes at targets numbered 1 through 5, spaced 250mm apart along a line, with the center 1m in front of the viewer. In our remote configuration, the middle of these targets, number 3, is co-located with a webcam. The webcam captures images of a person wearing the headset with no external display, the 2D display, the light field display, and for ground truth comparison, wearing no headset at all. These images, totaling 200 individual conditions, were then presented to 24 participants via a web survey and asked to select which of the 5 positions the headset-wearer was looking for each condition. The resulting response percentages are shown as histogram distributions in Figure 12.

In the center position (Figure 12c), we see that participants detected the correct at-camera gaze direction more frequently than the 2D or blank display cases. This is an important condition as it is a proxy for eye contact. Moving to the far left (Figure 12a) and far right (Figure 12e) cases, we see the same trend where the light field display outperforms everything but the no-headset case. The two intermediate gaze positions (Figure 12b and Figure 12d) are less clear. Note that the no-headset case is relatively poor for these two gaze positions, revealing the limitations of using uncalibrated 2D viewing setups in detecting subtle differences in gaze. We also see that many participants mistook gaze position 4 for 3 when displayed with light fields, while the no-headset and 2D cases were more often mistaken for position 5. Furthermore, the blank display results are somewhat biased toward the left, which we attributed to the uncontrolled but varying position of the wearer's head throughout the dataset capture. Other possible confounding factors include participant screen distance and size, which we could not control due to remote work requirements.

These limitations notwithstanding, our preliminary results support the conclusion in Section 5.1 that gaze is more accurately reproduced by light field displays. We performed one-way analysis of variance (ANOVA) on each target gaze position (resulting p-value listed in Table 1), which shows that users respond similarly to our light field displays and the no-headset case with only one gaze position having a statistically significant difference. For the 2D case, on the other hand, users responded with a statistically significant difference to the no-headset case for 3 of the 5 gaze targets. Because this study does not capture the autostereoscopic capability of the light field displays, which we believe will give our proposed technique a stronger advantage over 2D approaches, we hope to perform an in-person user study as soon as we can do so safely.

6 DISCUSSION

While our prototype has showcased our concept, there remain technical limitations that we anticipate will be pursued in future work.

Modeling Gaze Errors. The model described in Appendix B makes some assumptions, including a spherical eyeball with precisely known position and orientation, no refractive cornea, and pinhole projection. Moreover, the model makes no attempt to incorporate perceptual effects; it assumes projected gaze angle as a proxy for the perceived gaze angle. We believe these approximations are sufficient to perform quantitative comparisons between display methods.

In-Person User Studies. Since COVID restrictions prevented in-person studies involving our prototype, we were unable to conduct the most revealing mode of evaluation: a live user study where participants are able to experience the autostereoscopic display capabilities. We believe there are many studies to be conducted in the future around task performance-oriented metrics to evaluate whether people can reliably interpret the gaze of the headset wearer.

The introduction of an additional capture and display path will necessarily add costs to the headset, so user experience research will be necessary to determine whether those tradeoffs make the inclusion of reverse pass-through worthwhile.

Facial Reconstruction. We reiterate that the online facial reconstruction used in our experiments is only one example, designed around currently available hardware and software. We trained on a small dataset that does not generalize effectively. Ongoing research into the topic of view synthesis is likely to produce higher quality real-time facial models. Models optimized for telepresence, such as the one described by Lombardi et al. [2018], could generate facial reconstructions for both the local reverse pass-through view and remote VR views of the headset-wearer.

Improving Light Field Resolution and Field of View. As established in Section 3.2, thinner headsets will enable higher-resolution light field output due to the reduced distance between the external display and the user's face. Currently, holographic pancake architectures promise to approach sunglasses-like form-factors. Increased display density will also improve light field resolution by increasing the space-bandwidth product of the system. The increasing VR user base is is already driving the market toward high-density displays [Kopin 2021], though physical limits on pixel size remain.

Another consequence of ever-thinner headsets will be wider field of view requirements for the external display. Further development of microlens array optics will be necessary for these devices to be indistinguishable from a pair of glasses. Some system-level implementations can also effectively increase the field of view of current microlens array designs. For example, if outsider observers are sufficiently separated, their eye positions could be tracked, and then the relevant regions of the integral image optimally rendered for that position similar to Jones et al. [2014].

Benefits Relative to See-Through Augmented Reality. If these limitations can be overcome so that reverse pass-through systems are functionally equivalent to see-through AR displaya for outside observers, then VR devices may be preferable overall. While AR display resolution, field of view, accommodation support, and other characteristics may eventually be equivalent to blocked-light VR architectures, the fundamental see-through AR limitation of display contrast remains. AR displays additively combine with background content, meaning that even in the dimmest environments VR devices can reproduce higher fidelity virtual content. Some measures may be taken to mitigate this contrast loss, such as the placement of secondary attenuating displays in the optical path between the AR display and the real world. Unless novel angular-selective display technologies are developed, these attenuators will either require bulky relay optics (Wilson and Hua [2021]) or be limited to blurry regional dimming (Maimone et al. [2014]). On the other hand, digital compositing of virtual and real content in VR headsets has no such limitations, and can function in any lighting environment. Consequently, glasses form-factor VR headsets with reverse pass-through could ultimately provide better image quality than AR glasses in a wider range of environments.

7 CONCLUSION

The adoption of virtual reality in collaborative environments will require future headsets to support asymmetric interactions between headset-wearing and non-wearing participants. Today, headset-wearers can already see other people in their environment using video pass-through technologies. In our prior work, we designed and built the first system to use autostereoscopic world-facing display to enable realistic, multi-viewer social copresence in VR. In this paper, we detail the design and operation of reverse pass-through, illustrating the improvement in apparent gaze angle over prior 2D world-facing displays. We have supported this conclusion through a quantitative gaze error evaluation and a user study. In closing, we emphasize that reverse pass-through headsets have the potential to serve the same general-purpose use cases as optical see-through augmented reality glasses, including in social and professional copresence scenarios, while providing higher quality imagery to the wearer.

ACKNOWLEDGMENTS

We thank Tracy Campbell, Bruce Cleary, Phillip Guan, Salah Nouri, Sam Pepperwood, and Daniel Taylor for their contributions.

REFERENCES

Samer Al Moubayed, Jonas Beskow, Gabriel Skantze, and Björn Granström. 2012. Furhat: A Back-projected Human-like Robot Head for Multiparty Human-machine Interaction. Springer.

Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. 2009. The Digital Emily Project: Photoreal Facial Modeling and Animation.

Stuart M. Anstis, John W. Mayhew, and Tania Morley. 1969. The Perception of Where a Face or Television 'Portrait' is Looking. *American Journal of Psychology* 82, 4 (1969).

Amit H. Bermano, Markus Billeter, Daisuke Iwai, and Anselm Grundhöfer. 2017. Makeup Lamps: Live Augmentation of Human Faces via Projection. *Computer Graphics Forum* 36, 2.

Blender Foundation. 2021. *Blender - A 3D Modelling and Rendering Package*. Blender Foundation. http://www.blender.org

Liwei Chan and Kouta Minamizawa. 2017. FrontFace: Facilitating Communication between HMD Users and Outsiders Using Front-Facing-Screen HMDs. *ACM Human-Computer Interaction with Mobile Devices and Services*.

Milton Chen. 2002. Leveraging the Asymmetric Sensitivity of Eye Contact for Videoconference. *ACM Human Factors in Computing Systems*.

Sebastian Zambal Christoph Heindl, Lukas Brunner and Josef Scharinger. 2020. BlendTorch: A Real-Time, Adaptive Domain Randomization Library. *International Conference on Pattern Recognition Workshops*.

Christian Frueh, Avneesh Sud, and Vivek Kwatra. 2017. Headset Removal for Virtual and Mixed Reality. *ACM SIGGRAPH Talks* (2017).

Taichi Furukawa, Daisuke Yamamoto, Moe Sugawa, Roshan Peiris, and Kouta Minamizawa. 2019. TeleSight: Enabling Asymmetric Collaboration in VR between HMD User and Non-HMD Users.

Ying Geng, Jacques Gollier, Brian Wheelwright, Fenglin Peng, Yusufu Sulai, Brant Lewis, Ning Chan, Wai Sze Tiffany Lam, Alexander Fix, Douglas Lanman, et al. 2018. Viewing Optics for Immersive Near-eye Displays: Pupil Swim/Size and Weight/Stray Light. *Digital Optics for Immersive Displays*.

Daniel Gotsch, Xujing Zhang, Timothy Merritt, and Roel Vertegaal. 2018. TeleHuman2: A Cylindrical Light Field Teleconferencing System for Life-size 3D Human Telepresence. *ACM Conference on Human Factors in Computing Systems* 18.

Jan Gugenheimer, Christian Mai, Mark McGill, Julie Williamson, Frank Steinicke, and Ken Perlin. 2019. Challenges Using Head-Mounted Displays in Shared and Social Spaces. *ACM Human Factors in Computing Systems Extend Abstracts*.

Pieter Hintjens. 2013. *ZeroMQ: Messaging for Many Applications*. O'Reilly Media, Inc.

Fu-Chung Huang, David P Luebke, and Gordon Wetzstein. 2015. The Light Field Stereoscope. *ACM SIGGRAPH Emerging Technologies*.

Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving Eye Contact in a One-To-Many 3D Video Teleconferencing System. *ACM Transactions on Graphics* 28, 3 (2009).

Andrew V Jones, Koki Nagano, Jing Liu, Jay Busch, Xueming Yu, Mark T Bolas, and Paul Debevec. 2014. Interpolating Vertical Parallax for an Autostereoscopic Three-dimensional Projector Array. *Journal of Electronic Imaging* 23, 1 (2014).

Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. TeleHuman: Effects of 3D Perspective on Gaze and Pose Estimation with a Life-Size Cylindrical Telepresence Pod. *ACM Human Factors in Computing Systems*.

Kopin. 2021. Kopin's 2.6k x 2.6k OLED Display Incorporated in Panasonic's New VR Glasses. (Jan 2021). https://kopin.irpass.com/profiles/investor/NewsPrint.asp?v=6&b=2379&ID=96490&m=rl&g=1207

Douglas Lanman and David Luebke. 2013. Near-eye Light Field Displays. *ACM Transactions on Graphics* 32, 6 (2013).

Douglas Lanman, Gordon Wetzstein, Matthew Hirsch, and Ramesh Raskar. 2013. Depth of Field Analysis for Multilayer Automultiscopic Displays. *Journal of Physics: Conference Series* 415, 1.

Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. 2015. Facial Performance Sensing Head-Mounted Display. *ACM Transactions on Graphics* 34, 4 (2015).

Peter Lincoln, Andrew Nashel, Adrian Ilie, Herman Towles, Gregory Welch, and Henry Fuchs. 2009a. Multi-View Lenticular Display for Group Teleconferencing. *Immersive Telecommunications*.

Peter Lincoln, Greg Welch, Andrew Nashel, Adrian Ilie, Andrei State, and Henry Fuchs. 2009b. Animatronic Shader Lamps Avatars. *IEEE International Symposium on Mixed and Augmented Reality*.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Transactions on Graphics* 37, 4 (2018).

Christian Mai and Mohamed Khamis. 2018. Public HMDs: Modeling and Understanding User Behavior around Public Head-Mounted Displays. *ACM Pervasive Displays*.

Christian Mai, Lukas Rambold, and Mohamed Khamis. 2017. TransparentHMD: Revealing the HMD User's Face to Bystanders. *ACM Mobile and Ubiquitous Multimedia*.

Andrew Maimone, Douglas Lanman, Kishore Rathinavel, Kurtis Keller, David Luebke, and Henry Fuchs. 2014. Pinlight Displays: Wide Field Of View Augmented Reality Eyeglasses Using Defocused Point Light Sources.

Andrew Maimone and Junren Wang. 2020. Holographic Optics for Thin and Lightweight Virtual Reality. *ACM Transactions on Graphics* 39, 4 (2020).

Manuel Martínez-Corral and Bahram Javidi. 2018. Fundamentals of 3D Imaging and Displays: A Tutorial on Integral Imaging, Light-field, and Plenoptic Systems. *Advances in Optics and Photonics* 10, 3 (2018).

Nathan Matsuda, Brian Wheelwright, Joel Hegland, and Douglas Lanman. 2021. Reverse Pass-Through VR. *ACM SIGGRAPH Emerging Technologies*.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. *IEEE Computer Vision and Pattern Recognition*.

Kana Misawa and Jun Rekimoto. 2015. ChameleonMask: Embodied Physical and Social Telepresence using Human Surrogates. *ACM Human Factors in Computing Systems*.

Samer Al Moubayed, Jens Edlund, and Jonas Beskow. 2012. Taming Mona Lisa: Communicating Gaze Faithfully in 2D and 3D Facial Projections. *ACM Transactions on Interactive Intelligent Systems* 1, 2 (2012).

Koki Nagano, Andrew Jones, Jing Liu, Jay Busch, Xueming Yu, Mark Bolas, and Paul Debevec. 2013. An Autostereoscopic Projector Array Optimized for 3D Facial Display.

David Nguyen and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. *ACM Human Factors in Computing Systems*.

Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity Facial and Speech Animation for VR HMDs. *ACM Transactions on Graphics* 35, 6 (2016).

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. *User Interface Software and Technology*.

Colin J. Palmer, Yumiko Otsuka, and Colin W.G. Clifford. 2020. A sparkle in the eye: Illumination cues and lightness constancy in the perception of eye contact. *Cognition* 205 (2020).

Ye Pan and Anthony Steed. 2014. A Gaze-Preserving Situated Multiview Telepresence System. *ACM Human Factors in Computing Systems*.

Ye Pan and Anthony Steed. 2016. Effects of 3D perspective on Head Gaze Estimation with a Multiview Autostereoscopic Display. *International Journal of Human-Computer Studies* 86 (2016).

Ryan Schubert, Greg Welch, Peter Lincoln, Arjun Nagendran, Remo Pillat, and Henry Fuchs. 2012. Advances in Shader Lamps Avatars for Telepresence. *IEEE 3DTV*.

Valentin Schwind, Jens Reinhardt, Rufat Rzayev, Niels Henze, and Katrin Wolf. 2018. Virtual Reality on the Go? A Study on Social Acceptance of VR Glasses. *ACM Human-Computer Interaction with Mobile Devices and Services*.

David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. *IFIP Conference on Human-Computer Interaction*.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2018. FaceVR: Real-time Gaze-aware Facial Reenactment in Virtual Reality. *ACM Transactions on Graphics* 37, 2 (2018).

Chiu-Hsuan Wang, Seraphina Yong, Hsin-Yu Chen, Yuan-Syun Ye, and Liwei Chan. 2020. HMD Light: Sharing In-VR Experience via Head-mounted Projector for Asymmetric Interaction. *User Interface Software and Technology*.

Miao Wang, Xin Wen, and Shi-Min Hu. 2019b. Faithful Face Image Completion for HMD Occlusion Removal. *IEEE International Symposium on Mixed and Augmented Reality Adjunct*.

Yan Wang, Zihang Lai, Gao Huang, Brian H Wang, Laurens Van Der Maaten, Mark Campbell, and Kilian Q Weinberger. 2019a. Anytime Stereo Image Depth Estimation on Mobile Devices. *IEEE International Conference on Robotics and Automation*.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Transactions on Graphics* 38, 4 (2019).

Gordon Wetzstein, Wolfgang Heidrich, and Ramesh Raskar. 2014. Computational Schlieren Photography with Light Field Probes. *International Journal of Computer Vision* 110, 2 (2014).

Austin Wilson and Hong Hua. 2021. Design of a Pupil-Matched Occlusion-Capable Optical See-Through Wearable Display. *IEEE TVCG* (2021).

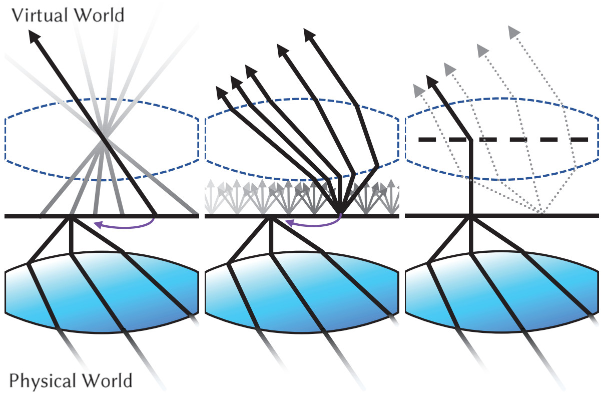
William Hyde Wollaston. 1824. On the Apparent Direction of Eyes in a Portrait. *Philosophical Transactions of the Royal Society of London* 114 (1824).

Timothy L Wong, Zhisheng Yun, Gregg Ambur, and Jo Etter. 2017. Folded Optics with Birefringent Reflective Polarizers. *Digital Optical Technologies*.

Chris Wyman. 2005. Interactive Image-space Refraction of Nearby Geometry. *International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*.

Tatsuo Yotsukura, Frank Nielsen, Kim Binsted, Shigeo Morishima, and Claudio S. Pinhanez. 2002. HyperMask: Talking Head Projected onto Real Object. *The Visual Computer* 18, 2 (2002).

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-image Translation Using Cycle-consistent Adversarial Networks. *IEEE International Conference on Computer Vision*.



(a) Camera-Array (b) Full Raytracing (c) Thin Refractor

Fig. 13. Generating light field display integral images is commonly achieved by rasterizing a pinhole camera view for each lenslet subview (a). The pixel coordinates of the resulting image must be flipped about each lenslet center to rectify the output light field (purple arrow). Rasterizing a scene using thousands of unique projection matrices poses performance challenges. Tracing rays emitted from each pixel location on the display plane through an accurate refractive model of the microlens array (b) can capture more complex effects of the physical system, but is even more challenging for real-time operation. We find an ideal thin refractor that deflects an orthogonal ray to an outgoing ray (thick black arrow) that matches the output ray directions for an offline high quality raytrace (thin dotted arrows) through an optically-correct model of the microlens array, including the rectification step. Rendering a high resolution integral image with this technique can be accomplished at high framerates using conventional screen-space refraction models.

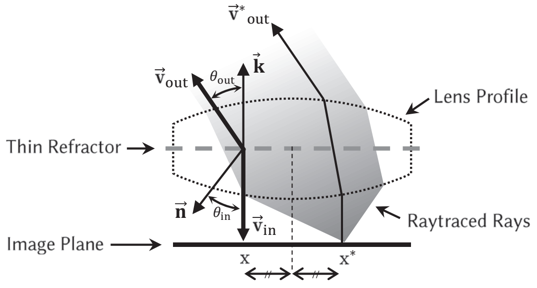


Fig. 14. Raytracing an accurate lens profile, modeled after the manufactured microlens prescription, produces a target outgoing ray direction $\vec{\mathbf{v}}^*_{\mathrm{out}}$ for each position x^* at the display plane. We then solve for the normal map $\vec{\mathbf{n}}$ of a thin refractor that produces an outgoing ray $\vec{\mathbf{v}}_{\mathrm{out}}$ parallel to the target direction given an incoming ray $\vec{\mathbf{v}}_{\mathrm{in}}$ that originates at the inverted screen position x normal to the display plane. This last inversion accounts for the sub-image flipping during projection through the microlens during playback. The normal map is tiled to the full microlens array pattern and passed to a screen-space refraction shader. Images to be displayed on the physical light field display can then be rendered in one shot with an orthographic camera matching the size and resolution of the physical display.

A LIGHT FIELD RENDERING

The ideal light field display directly addresses intensity at discrete spatial and angular coordinates at the plane of the display. In contrast, a physical microlens array (MLA) allows the device to approximate such functionality by imaging small portions of a 2D LCD to a distant focal surface (which is not precisely a plane nor exactly at infinity, due to aberrations and small misalignments). In order to operate our display, we need to bridge these two concepts—efficiently mapping ray angles to positions on the LCD display, while accounting for the optical properties of the microlenses.

A common approach to producing a simple mapping from screen position to ray angle is depicted in Figure 13a. An array of cameras is constructed such that the center of projection for each microlens has a corresponding render camera. These cameras need only sample the number of pixels spanned by the microlens, which is typically a very small number. However, this approach can incur high rendering overhead due to the need for a unique projection matrix to be loaded for each camera, which can easily number in the thousands (our small 2" display has around $5,000$ sub-views). Additionally, if the microlens array has a non-gridded packing (such as our hex-packed MLA), the rasterization for each camera must be resampled into the correct output pixel position, a step that incurs a nontrivial amount of additional compute and memory cost.

Figure 13b depicts a physically accurate, "brute force" alternative rendering method: a raytrace is initiated with ray sampling covering an image plane corresponding to the physical display plane. These rays are then refracted through a physically correct model of the microlens array, and then out into the virtual scene. While this can reproduce higher-order characteristics of the MLA, it requires precise calibration of the entire MLA, and of course it is very computationally expensive.

We break this problem down. First, we perform a forward raytrace through a geometric model of a single lenslet (but without a virtual scene), recording only the outgoing direction for each point (grey dotted lines in Figure 13c). Then we solve for an ideal thin refractor at the midplane of the lenslet (horizontal dashed line) that produces the same outgoing direction as the raytrace for a ray parallel to the optical axis and with the inverted origin (thick arrow), which is needed to account for the inverted image produced by the microlens during playback. This thin refraction map can then be tiled via texture mapping in whatever configuration the MLA happens to be. A standard screen-space refraction model [Wyman 2005] can then produce an accurate integral image for the entire MLA in one step (two passes), allowing for high framerates even at high display resolutions.

The thin refractor will be defined by an arbitrary refractive index n and a spatially varying normal map, where each position x, relative to the center of the lenslet, has an associated normal vector $\vec{\mathbf{n}}(x)$. An incoming ray $\vec{\mathbf{v}}_{\mathrm{in}}$ will be deflected to outgoing ray $\vec{\mathbf{v}}_{\mathrm{out}}$, which we want to be parallel to the optically correct outgoing ray $\vec{\mathbf{v}}^*_{\mathrm{out}}$ (See Figure 14).

To simplify the selection of $\vec{\mathbf{v}}_{\mathrm{in}}$ and enable the use of a standard camera model, we make the observation that for every position x^* on the plane imaged by the microlens, there is a ray parallel to the optical axis $\vec{\mathbf{k}}$ that intersects that point and then exits in the direction $\vec{\mathbf{v}}^*_{\text{out}}$ (thin black ray in Figure 14).

Following Wetzstein et al. [2014], the relative angle between the refractive normal and incoming orthogonal ray will be θ_{in}:

$$\cos\theta_{\text{d}} = \vec{\mathbf{k}} \cdot \vec{\mathbf{v}}_{\text{out}}$$
$$\theta_{\text{in}} = -\tan^{-1}\left(\frac{\sin(\theta_d)}{\cos(\theta_d) - \frac{1}{n}}\right) \tag{1}$$

Finally, we recover $\vec{\mathbf{n}}(x)$ directly by rotating $\vec{\mathbf{v}}_{\text{in}}$ by θ_{in} about the vector perpendicular to $\vec{\mathbf{v}}_{\text{in}}$ and $\vec{\mathbf{v}}_{\text{out}}$:

$$\vec{\mathbf{n}} = \mathbf{R}(-\theta_{\text{in}}, \vec{\mathbf{k}} \times \vec{\mathbf{v}}_{\text{out}}) \tag{2}$$

Since we precomputed $\vec{\mathbf{v}}_{\text{out}}(x^*)$, and $x = -x^*$ to rectify the subview, the normal map is a function of x alone. If we apply this normal map to a plane with a refraction shader and render with a frontoparallel orthographic camera, we will have mapped incoming angle to image location. Each pixel samples a single outgoing ray that falls within the collimated ray bundle passing through the lenslet. Conventional anti-aliasing achieved by jittering sample locations in the camera plane will result in diverging ray bundles, but we can instead jitter the camera position and MLA refraction plane together to approximate the integral over the collimated ray bundles.

We can gain additional computational efficiency by using screen-space refraction [Wyman 2005], which stores an RGB-D image of the scene in the first pass and then retrieves refracted ray color values from this buffer in a second pass. This has the drawback of eliminating view-dependent shading in the light field image as the RGB-D image discards angular information. However, using true raytraced refraction would retain these effects.

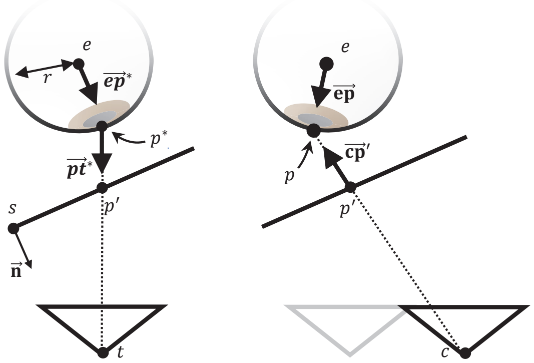
We implement this approach in Python and Blender, using the latter's physically based Cycles engine for the precomputed raytrace, which is stored as an EXR texture map, and the built-in EEVEE engine's screen-space refraction feature for efficient run-time playback. To support refraction rays that terminate outside the bounds of first render pass (a fundamental limitation of screen-space refraction), we overscan the render buffer to cover the full area of the reconstructed face that can be addressed by the MLA.

All results in this paper, both simulated and experimental captures, use this method for generating light field images.

B APPROXIMATING APPARENT GAZE ERROR

Several related publications listed in Section 2 use a 2D screen that displays a reprojected image based on a tracked estimate of the viewer's position so that perspective appears correct. Because of the lack of autostereoscopic output, fidelity of these tracked 2D displays will be fundamentally limited by the distance between the display and virtual image, the distance to the viewer, and the interpupillary distance of the viewer. Tracking errors will also degrade the quality of the perspective correction, but even with perfect tracking there is a limit to the accuracy of apparent gaze direction of an eye shown on such a display. Projection accuracy will, in turn, limit the accuracy of the perceived gaze direction.

To quantify this limit, we use a reprojection model depicted in Figure 15. In Figure 15a, the eye position e, true gaze direction $\overrightarrow{\mathbf{ep}}^*$, and eye radius r are used to produce true pupil position p^*. This pupil position is projected to the plane of the display screen (defined by reference point s and normal $\vec{\mathbf{n}}$), given tracked view position t:



(a) Target Projection (b) Apparent Projection

Fig. 15. We estimate apparent gaze angle error for 2D tracked approaches by projecting the true pupil position p^* to its position p' on the screen plane defined by position s and normal $\vec{\mathbf{n}}$ from tracked view position t. Then, in (b), we find the apparent pupil position p by projecting p' back onto the eyeball. We then compare apparent gaze direction $\overrightarrow{\mathbf{ep}}$ to true direction $\overrightarrow{\mathbf{ep}}^*$. We make the assumption that the forward and backward projection do not change the apparent position of the eyeball since the eyeball outline is not visible to the viewer and is instead inferred from the head position.

$$\begin{aligned} \overrightarrow{\mathbf{pt}}^* &= \frac{p^* - t}{\|p^* - t\|} \\ d &= \frac{(s-t)\cdot\vec{\mathbf{n}}}{\overrightarrow{\mathbf{pt}}^* \cdot \vec{\mathbf{n}}} \\ p' &= t + d\overrightarrow{\mathbf{pt}}^* \end{aligned} \tag{3}$$

Now we project p' back to the eyeball from the point along the ray $\overrightarrow{\mathbf{cp}}'$, originating at the actual view center c, to find the apparent pupil position p:

$$\begin{aligned} \overrightarrow{\mathbf{cp}}' &= \frac{p' - c}{\|p' - c\|} \\ \Delta = (\overrightarrow{\mathbf{cp}}' \cdot (p' - e))^2 &- (\|p' - e\|^2 - r^2) \\ d' &= -(\overrightarrow{\mathbf{cp}}' \cdot (p' - e)) - \sqrt{\Delta} \\ p &= p' + d'\overrightarrow{\mathbf{cp}}' \end{aligned} \tag{4}$$

We can compare $\overrightarrow{\mathbf{ep}} = \frac{p-e}{\|p-e\|}$ to $\overrightarrow{\mathbf{ep}}^*$ and approximate the gaze error. Because the eyeball geometry itself cannot be seen by the viewer in a real scenario, we make the assumption here that the apparent eyeball position is inferred by the viewer given the rest of the facial geometry. While we acknowledge this avoids complex human perception and expectations about facial and eye geometry, the simplification enables a useful numerical comparison of gaze-preserving social copresence systems.