Perspective-Correct VR Passthrough Without Reprojection

Grace Kuo
Reality Labs Research, Meta
USA
gracekuo@meta.com

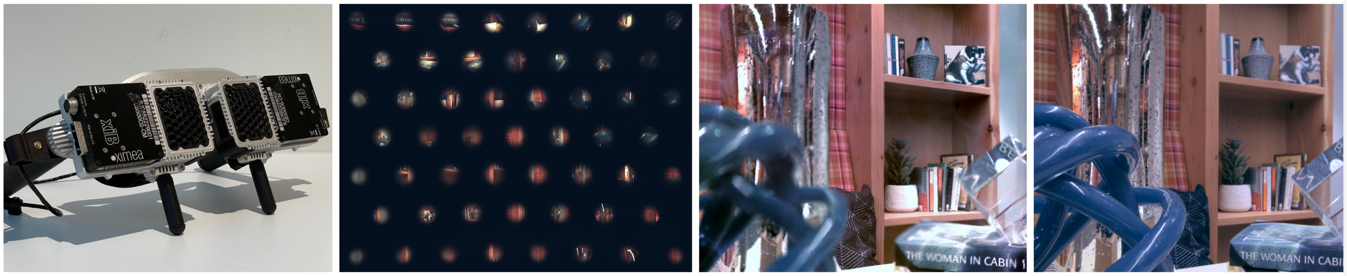
Eric Penner
Reality Labs Research, Meta
USA
epenner@meta.com

Seth Moczydlowski
Reality Labs Research, Meta
USA
smoczydlowski@meta.com

Alexander Ching
Reality Labs Research, Meta
USA
alexching@meta.com

Douglas Lanman
Reality Labs Research, Meta
USA
douglas.lanman@meta.com

Nathan Matsuda
Reality Labs Research, Meta
USA
nathan.matsuda@meta.com



(a) Prototype passthrough headset (b) Raw sensor data (c) Reconstructed view at eye (d) Reference camera at eye

Figure 1: Virtual reality (VR) passthrough allows VR users to interact with their environment via video streams from external cameras. However, passthrough cameras on the front of the headset capture a different perspective than what the user would see natively. To address this, we present a computational camera based on a modified lens array that captures a perspective from *behind* the the sensor, enabling direct capture of the view at the user's eye in a compact form factor. We demonstrate this concept with a working prototype headset and a low latency algorithm that faithfully reconstructs the perspective at the eye.

ABSTRACT

Virtual reality (VR) passthrough uses external cameras on the front of a headset to allow the user to see their environment. However, passthrough cameras cannot physically be co-located with the user's eyes, so the passthrough images have a different perspective than what the user would see without the headset. Although the images can be computationally reprojected into the desired view, errors in depth estimation, view-dependent effects, and missing information at occlusion boundaries can lead to undesirable artifacts.

We propose a novel computational camera that directly samples the rays that would have gone into the user's eye, several centimeters behind the sensor. Our design contains an array of lenses with an aperture behind each lens, and the apertures are strategically placed to allow through *only* the desired rays. The resulting thin, flat architecture has suitable form factor for VR, and the image reconstruction is computationally lightweight, enabling low-latency passthrough. We demonstrate our approach experimentally in a fully functional binocular passthrough prototype with practical calibration and real-time image reconstruction. Finally, we experimentally validate that our camera captures the correct perspective for VR passthrough, even in the presence of transparent objects, specular highlights, and complex occluding structures.

CCS CONCEPTS

• **Hardware → Displays and imagers**; • **Human-centered computing → Mixed / augmented reality**; **Virtual reality**.

KEYWORDS

virtual reality, passthrough, light field

ACM Reference Format:
Grace Kuo, Eric Penner, Seth Moczydlowski, Alexander Ching, Douglas Lanman, and Nathan Matsuda. 2023. Perspective-Correct VR Passthrough Without Reprojection. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 06–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3588432.3591534

1 INTRODUCTION

Virtual reality (VR) head-mounted displays (HMDs) offer immersive experiences thanks to their high contrast imagery and large field-of-view (FoV), but users are isolated in the virtual world since the

SIGGRAPH '23 Conference Proceedings, August 06–10, 2023, Los Angeles, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0159-7/23/08.
https://doi.org/10.1145/3588432.3591534

headset blocks the user from seeing their surroundings. In contrast, augmented reality (AR) uses transparent displays, allowing users to be fully present in their environment, but AR displays today have limited FoV. In addition, virtual content in AR must compete with environmental lighting, reducing display contrast.

VR passthrough offers a compromise. By streaming video from external facing cameras into a VR headset, passthrough allows users to view their surroundings without sacrificing the advantages of VR hardware. For users to seamlessly interact with their environment, the displayed passthrough image should match what the user would see without the headset. However, passthrough cameras located on the front of the headset capture a different perspective than the view from the user's eye. Even if passthrough cameras are placed directly in front of the eyes, there is still an axial offset due to the thickness of the HMD. Streaming these images directly to the user causes *visual displacement.*

Visual displacement can be corrected computationally by estimating the depth of objects in the scene and reprojecting the camera views to the eye location [Chaurasia et al. 2020; Xiao et al. 2022]. However, errors in the depth estimation and missing information at occlusion boundaries can create artifacts in the passthrough image. Furthermore, the view synthesis algorithm must run with low latency, potentially on mobile hardware.

We propose an alternate approach in which we design the optical hardware and a corresponding image reconstruction algorithm specifically for passthrough. Our novel camera design directly measures the exact rays that would have gone into the eye, located *behind* the camera sensor. Unlike prior architectures that optically capture the correct view with mirrors or prisms, our design is thin and flat, allowing it to meet the form factor requirements of VR. We refer to our approach as *light field passthrough.*

We make the following contributions:

- A novel computational camera for VR passthrough based on our light field passthrough architecture, which perfectly captures images from the perspective of a virtual eye behind the camera in a compact form factor suitable for HMDs.
- Analysis of the design space of light field passthrough.
- A practical calibration technique and computationally lightweight algorithm for image reconstruction, including coarse depth estimation and gradient domain image stitching, with combined runtime under 1.7 ms per frame.
- Design and demonstration of a fully functional binocular passthrough HMD, and experimental validation that our approach captures the correct perspective, even for challenging scenes with near-field content, specular reflections, and transparent objects.

2 RELATED WORK

Effects of Visual Displacement in Video Passthrough. Visual displacement between passthrough cameras and the user's eyes can cause negative perceptual effects and make it more challenging for users to interact with the world. In the first study on the topic, Rolland et al. [1995] built a video passthrough system with axial and vertical displacement of the cameras (165 mm and 62 mm, respectively) and found that users were slower by 43% at manual tasks and had significant pointing errors compared to a transparent headset. Although pointing errors reduced as the users adapted to the visual displacement, they never returned to baseline, and participants experienced negative after-effects in the form of increased errors after the headset was removed [Biocca and Rolland 1998]. Park et al. [2008] followed up on this work, comparing hand-eye coordination in headsets over a range of visual displacements, including using a mirror to place the cameras at the user's eye. They found that users performed tracing tasks faster with the mirror configuration, compared to when the cameras where located in front of the headset, even with no vertical displacement. However, participants were faster and more accurate without the headset at all, suggesting visual displacement is not the only characteristic of VR passthrough affecting task performance. In another study, Lee et al. [2013] tested adaptation to visual displacement over several configurations, and found participants adapted within 10 minutes. However, participants also reported a feeling of "body structure distortion" in which they had the sensation that their limbs were attached to their bodies in the wrong locations. Studies of large lateral displacements (50 mm - 300 mm) found that task performance decreased with larger visual displacement [Lee and Park 2020] and simulator sickness increased [Kim et al. 2014].

In contrast to the other studies, Takagi et al. [2000] found users could estimate the size and position of objects equally well when the cameras were at the eye versus when they were displaced up to 40 mm axially. Although this suggests that modest displacements may be acceptable, we point out that the resolution of the study's headset was only 640 × 480, and modern displays have over an order of magnitude more pixels. More recent work by Guan et al. [2023] showed that users are sensitive to axial displacement of only 15 mm, which is shorter than the eye relief of glasses. Krajancich et al. [2020] found that, in some scenarios, users can detect millimeters of displacement due to eye rotation.

Computational Passthrough. One option to remove visual displacement is to computationally synthesize the view at the eye location from a small number of cameras on the front of the headset. View synthesis is a widely-researched problem; popular solutions include classical image-based rendering [Chen and Williams 1993], multiplane images [Zhou et al. 2018], and neural radiance fields [Mildenhall et al. 2021]. However, these techniques are not tailored for VR passthrough, which requires low latency and temporally consistent view synthesis. To address this gap, Schöps et al. [2017] demonstrated real-time view synthesis using edge-aware inpainting given a precomputed depth map, and Chaurasia et al. [2020] proposed an end-to-end passthrough algorithm, commercially available on Meta products such as Quest 2, in which depth is estimated from a stereo pair of cameras. Although this work enables real-time passthrough on mobile devices, there can be significant warping artifacts from inaccuracies in the depth estimation, particularly around near objects and occlusion boundaries. More recently, Xiao et al. [2022] proposed a neural passthrough approach using modern machine learning techniques to improve depth estimation and fill in missing information from occlusions. However, specularities and repeating patterns in the scene can cause artifacts, and the algorithm is too computationally intensive for mobile headsets.

Optical Architectures for Perspective-Correct Passthrough. Another option for removing visual displacement is to design an optical

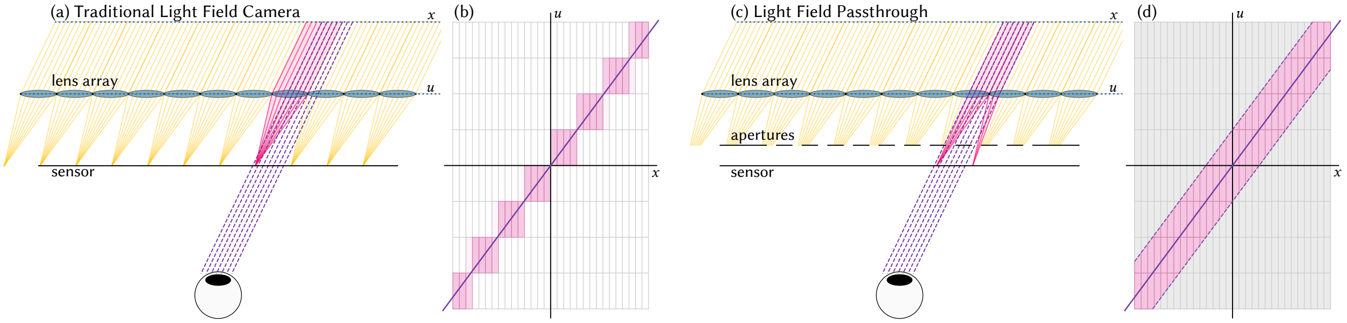

(a) Traditional Light Field Camera
x
lens array
u
sensor
(b)
u
x
(c) Light Field Passthrough
x
lens array
u
apertures
sensor
(d)
u
x

Figure 2: Passthrough aims to synthesize a new view at the user's eye, some distance behind the camera. (a) A traditional light field camera uses an array of lenses to capture the scene from a grid of different view-points. To form a new view, the light field can be resampled, depicted in epipolar space in (b) by selecting the ray bundles (pink) corresponding to the target view represented by the purple line. However, the coarse resolution of the light field results in inaccurate sampling of the rays and discontinuities in the synthesized view. (c) We overcome these issues with our light field passthrough architecture in which we add an aperture behind each lens; the apertures are designed to physically block rays that would not enter the eye. In epipolar space (d), only the ray bundle highlighted in pink reaches the sensor, enabling theoretically perfect reconstruction of the target view.

architecture that directly captures the rays that would have gone into the eye, therefore capturing the correct perspective. In these architectures, the camera view doesn't require post processing beyond distortion correction and can be streamed directly to the user with low latency. Edwards et al. [1993] describe how this can be accomplished with a mirror at a 45° angle in front of the headset, which folds the optical path to a camera placed above or below the device; Fuchs et al. [1998] built a functional version of this design. However, the form factor cost is high as the mirror sticks out, approximately doubling the headset track length. Takagi et al. [2000] replaced the mirror with a prism, which uses total internal reflection (TIR) to fold the light path, reducing camera form factor. However, the solid mass of the prism adds substantial weight, and the FoV will be limited to angles that match the TIR condition.

Light Field Cameras. A dense light field contains all the information necessary to synthesize novel views [Levoy and Hanrahan 1996; Ng et al. 2005], making light fields a temping candidate for passthrough. However, dense light fields are traditionally captured by scanning the camera location [Levoy et al. 2000], which is unsuitable for real-time applications, or using large camera arrays [Broxton et al. 2020], which is unrealistic for HMDs. Although compact, single-exposure versions have been described using lens arrays [Tanida et al. 2001], angular sensitive pixels [Wang et al. 2009], Frensel zone plates [Shimano et al. 2018], and metasurfaces [Lin et al. 2019], light fields have a trade-off between spatial and angular resolution. Sub-pixel sampling [Georgiev et al. 2011] can improve resolution by a factor of two, but the trade-off still exists. Compressive light field techniques [Marwah et al. 2013] that overcome the trade-off have prohibitively high computational cost for passthrough. As a result, existing real-time devices either have low spatial resolution or insufficient angular sampling for artifact-free view synthesis.

3 SYSTEM OVERVIEW

We propose a passthrough camera architecture that optically captures the correct perspective, similar to a dense light field or a mirror based system, while maintaining the practical form factor of a lens array. Our design is inspired by view synthesis in light fields, summarized below.

3.1 Traditional Light Field View Synthesis

A light field camera, depicted in Fig. 2a, a uses an array of lenses in front of a sensor to capture the intensity of light as a function of both position and angle of the incoming rays. As described by Ng et al. [2005], we can parameterize the rays by their intersection at two planes, denoted x and u, and plot the location of rays in epipolar space, shown in Fig. 2b. Each box represents the bundle of rays captured by a single pixel on the detector.

For video passthrough, we would like to synthesize a new view at the user's eye location, some distance behind the camera, where the distance is determined by the thickness of the HMD. We can form this view from the light field by sampling the rays that correspond to the view of interest, denoted by the purple line in Fig. 2b. With an infinitely high resolution light field, we could perfectly form the desired view. However, the resolution is limited by the number of pixels on the sensor, which must be distributed over the 4D light field. This results in a trade-off between angular and spatial resolution, both of which must be high to synthesize a new high resolution view. Furthermore, the angular resolution (along u) is determined by the size of each lens. Smaller lenses result in higher angular resolution; however, as lens diameter shrinks, diffraction starts to reduce the spatial resolution.

With a limited resolution light field, we can select the ray bundles that best correspond to the view of interest (highlighted in pink in Fig. 2b). However, the resulting view is an approximation; as shown in Fig. 2a, the selected ray bundle is different from the rays that would have gone into the user's eye without the headset. In fact, there is only one angle per lens where the selected rays match the desired rays. The most challenging areas are at the boundaries between lenses, where the synthesized view may have distracting boundary artifacts if there is more than 1 pixel of disparity between views. See the supplemental video for a visualization for this effect.

3.2 Light Field Passthrough

We propose a modification to the traditional light field camera, tailoring it for video passthrough in which only a single novel view needs to be generated. Our key design innovation is to add a carefully designed array of apertures, one behind each lens. The apertures physically block all of the rays that would not have entered the user’s eye (Fig. 2c). In epipolar space (Fig. 2d), the apertures block the rays shown in gray, leaving only rays around the purple line, which represents the desired view. Importantly, the apertures don’t just remove samples, they also change the shape of the measurements in epipolar space, clipping them along the dotted lines in Fig. 2d. To understand why, notice in Fig. 2c how the pink rays that reach the sensor only cover a fraction of their corresponding lens; in comparison, in a traditional light field camera, each measured ray bundle always covers the whole area of the lens, resulting in rectangular samples in epipolar space.

Adding the aperture array enables measurement of the *exact* ray bundle that would have gone into the user’s eye; we’ll refer to this architecture as *light field passthrough.* In addition to capturing the correct view and creating seamless transitions between lenses, by blocking most of the light field, we can better distribute the finite number of pixels on the sensor. In Fig. 2b, one can see that most of the samples aren’t used to form the new view; in light field passthrough, we can concentrate the samples in the region of interest, resulting in higher spatial resolution in the final image.

4 SYSTEM DESIGN

4.1 Aperture Locations

The aperture placement is the key element of light field passthrough; we want the apertures to allow through only rays of light that would have gone into the eye, regardless of the incident angle. This happens when the entrance pupil of each lens is at the eye location. By definition, the entrance pupil is the image of the aperture as seen through the lens from the object side [Hecht 2012]. We define a virtual aperture at the eye, and for each lens in the array, its physical aperture and the virtual aperture at the eye should be conjugates.

Assuming ideal optics, we can use the thin lens equation to determine the locations of the physical apertures for a given virtual aperture (i.e. entrance pupil) position. If the entrance pupil is located z_{eye} behind the sensor, for a lens with focal length f, the associated physical aperture is at

$$z_{ap} = \frac{f^2}{2f + z_{eye}} \tag{1}$$

$$u_{ap} = \left(\frac{z_{eye} + z_{ap}}{z_{eye} + f} \right) u_{lens} \tag{2}$$

where z_{ap} is the vertical distance from the sensor to the physical aperture, and u_{lens}, u_{ap} are the lateral distances from the entrance pupil to the center of the lens and physical aperture, respectively. If the entrance pupil has diameter d_{eye}, then the physical aperture has diameter

$$d_{ap} = \left(\frac{f - z_{ap}}{f + z_{eye}} \right) d_{eye} \tag{3}$$

Here, we assumed the lens is focused at optical infinity (e.g. located f above the sensor). See Supplement for the equations when the lens is focused at different distances.

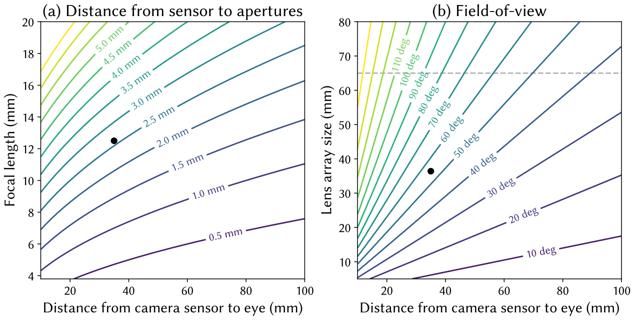


Figure 3: Thinner HMDs can reduce camera thickness and increase system FoV. (a) The lens focal length sets the camera thickness but must be chosen to accommodate the physical size of the apertures. (b) Larger FoV requires a physically larger lens array. However, the lens array size cannot exceed the user’s interpupillary distance (dashed line) in a binocular system. Black dots represent the parameters of our prototype, described in Sec. 6.1.

Since Eqs. (1) - (3) assume ideal optics, lens aberrations can change the relationship between the entrance pupil and physical aperture locations. Luckily, with ray tracing, we can calculate the aperture location accurately for any imaging lens. To do this, one first defines the entrance pupil shape and location. Starting with a single point on the boundary of the entrance pupil, trace the rays that would intersect that point. Once the rays pass through the lens, they will approximately converge at a new point behind the lens. Aberrations will generally cause some spread of the rays, but one can estimate the 3D spot where the rays converge; this is one edge of the physical aperture. By repeating this procedure for every point on the boundary of the entrance pupil, one can trace out the boundary of the corresponding physical aperture; in many cases, symmetry can be used to generate the aperture boundary by tracing only a small number of points.

However, even though the ray tracing algorithm is more accurate, the ideal lens equations (Eqs. (1) - (3)) can provide valuable insight into the design space of light field passthrough cameras, which we go into next.

4.2 Design Considerations

We describe a few design considerations unique to light field passthrough, and our analysis points toward thinner VR headsets for a better passthrough experience. Other design considerations on resolution, depth-of-field (DoF), uniformity, and redundancy are included in the Supplement.

Form Factor and Physical Constraints. The focal length of the lens array, f, approximately sets the thickness of the passthrough camera. Figure 3a and Eq. (1) show that choosing shorter focal lengths reduces z_{ap}, moving the apertures closer to the sensor. In practice, physical limitations may prevent the apertures from being arbitrarily close to the sensor; for example camera coverglass could limit aperture placement and the apertures themselves may have some non-negligible thickness. As the distance from the camera to the eye gets larger, one may need to choose a longer focal length lens to accommodate these restrictions. This suggests that better camera form factors will be achieved when the headset itself is also thin, enabling shorter z_{eye}.

Field-of-View. The FoV, plotted in Fig. 3b, is determined by the headset track length and the lateral size of the lens array. As z_{eye} increases with thicker headsets, we require larger lens arrays; if the lens array size exceeds the interpupillary distance of the user (dashed line in Fig. 3b), the lens arrays for each eye physically overlap which is not possible with our current architecture. It may be possible to design for this situation by placing two apertures per lens in the overlap region, but this type of design is outside the scope of this paper. Instead we note that VR is trending towards thinner headsets through the use of pancake lenses [Geng et al. 2018], such as in the Meta Quest Pro, and diffractive optics [Maimone and Wang 2020]. These thinner headsets, on the order of 30 mm from eye to front surface, can enable a 90° FoV without any physical overlap between the cameras associated with each eye.

5 CALIBRATION AND RECONSTRUCTION

Light field passthrough leverages the optical design to improve passthrough accuracy and reduce the computational burden. However, the raw sensor data (Fig. 1b) still requires post-processing as it consists of many sub-aperture views rather than a complete image. Theoretically, with an ideal lens array, we could use prior knowledge of the lens locations to simply rearrange the sub-aperture views into the final passthrough image. However, in practice, the exact lens locations may be unknown, and, furthermore, distortion within each lens can create discontinuities or ghosts artifacts in the reconstruction. We correct for both the unknown lens locations and the distortion simultaneously by calibrating a dense mapping from sensor pixels to output image pixels using Gray codes displayed on a television [Bitner et al. 1976; Sels et al. 2019]. Then, a basic reconstruction algorithm consists of rearranging the pixels based on the calibration and applying a flat-field correction. This algorithm extremely lightweight making it suitable for low latency passthrough. However, close inspection reveals ghost artifacts due to depth dependence of the calibration and visible seams between sub-images due to stray light. We now go into more detail on where these artifacts come from and how to correct them.

5.1 Depth-Dependent Reconstruction

To understand why the calibration is depth-dependent, consider the diagram in Fig. 4a which depicts the rays corresponding to two points at different depths. p_1, at the further depth, is in focus on the sensor, and p_2, at the closer depth, focuses behind the sensor, creating defocus blur in the raw data (Fig. 4b). If the calibration is done at the plane of p_1, the points in the raw data corresponding to p_1 will line up in the reconstruction; the result is shown in Fig. 4c, where p_1 is in focus and p_2 is defocused, as expected. Note that the defocus bokeh of p_2 is spread between the two lenses and does not get overlapped in the final reconstruction resulting in more defocus blur in the final image than in each sub-aperture view individually. In contrast if the calibration is done at the plane of p_2, the points in the raw data corresponding to p_2 are aligned, resulting in the reconstructed intensity of Fig. 4d, with the undesirable consequence of p_1 being doubled in the image.

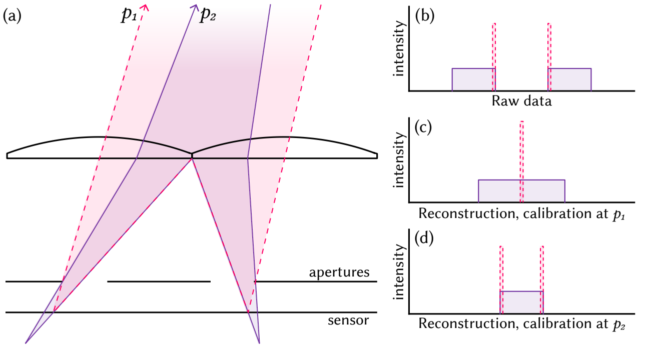

(a)
p_1
p_2
apertures
sensor
(b)
intensity
Raw data
(c)
intensity
Reconstruction, calibration at p_1
(d)
intensity
Reconstruction, calibration at p_2

Figure 4: Calibration works best at the focal plane. (a) Diagram depicting lenses focused a point p_1 (dashed). A closer point (p_2, solid) focuses behind the sensor, creating defocus blur in the raw data, shown in (b). If the system is calibrated at the focal plane (p_1), the resulting reconstruction does not have any doubling (c). However, if the system is calibrated off the focal plane (p_2), then the reconstruction contains undesirable doubling of in-focus content (d). However, in a real system, there may not be a single focal plane due to lens aberrations, in which case a depth-dependent reconstruction is necessary to remove doubling artifacts.

This example implies that the calibration should be done at the focal plane of the camera. However, due to field curvature of the lenses, in our prototype there is no single plane in where all the lenses are simultaneously in focus, resulting in doubling artifacts for objects off the plane of calibration (see Supplement for a schematic). This is confirmed in the experimental example in Fig. 5a,b depicting the reconstruction with calibration at optical infinity and 0.33 m, respectively. When calibration is at the background, there are artifacts in the foreground. These are resolved with a foreground calibration, but the background then has severe doubling. In this scenario, we need a depth-dependent reconstruction.

Given a depth map, we can reconstruct a clean image over the full FoV by choosing the correct calibration for each pixel. Since many mixed reality applications may require depth maps, one option is to simply leverage those. If that's not an option, we propose a coarse depth estimation using the overlap between lenses with an adapted block-matching approach. Although these depth maps are very low resolution and contain only a handful of planes, in Fig. 5c, the improvement using the depth estimation is apparent since both the foreground and background are reconstructed without doubling artifacts. We'd like to point out that, unlike traditional depth-based reprojection, inaccuracies in our depth map have minimal consequence on the reconstruction since points in the output only move

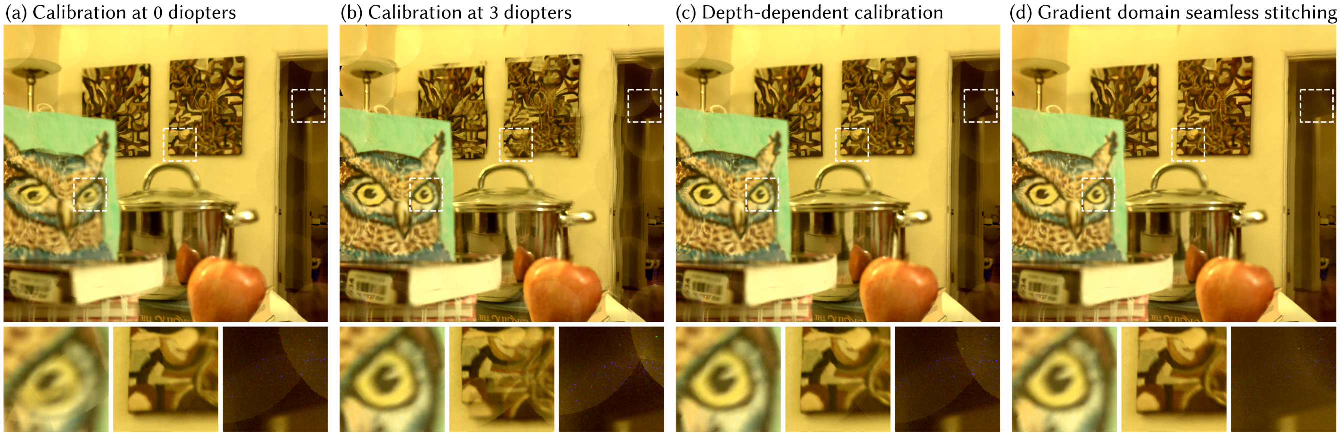

(a) Calibration at 0 diopters
(b) Calibration at 3 diopters
(c) Depth-dependent calibration
(d) Gradient domain seamless stitching

Figure 5: Depth-dependent calibration and gradient domain image stitching. A single calibration has artifacts on scenes with a large depth range. Calibration at optical infinity (a) results in artifacts in the foreground, and vice versa (b). We use the overlap between neighboring lenses to estimate a rough depth map, enabling a depth-dependent reconstruction (c) which removes artifacts over the whole FoV. Finally, we apply gradient domain image stitching (d) to remove visible boundaries between sub-aperture views and to get our final passthrough image.

by a small number of pixels as a function of depth. See Supplement for for more implementation details and example depth maps.

5.2 Gradient Domain Image Stitching

Even with a flat-field correction, stray light can cause differences in intensity between sub-aperture views resulting in noticeable seams in the reconstruction. We propose applying gradient domain image editing (GDIE) techniques for seamless stitching [Levin et al. 2004]. We compute spatial gradients of each sub-aperture image, smoothly blend them into a single gradient image, then convert back to the image domain using the FFT method of Frankot and Chellappa [1988], which can be implemented efficiently on a graphics processing unit (GPU). We optionally constrain the low frequencies of the gradient domain output to match those of the reconstruction without GDIE blending. Figure 5d shows an example reconstruction; note that the seams are almost entirely invisible with GDIE.

6 RESULTS

We demonstrate our light field passthrough concept with a binocular headset prototype (Fig. 1a). We implement all parts of the algorithm in real-time and validate that the camera captures the desired view.

6.1 Headset Design and Implementation

Based on Fig. 3b, to maximize the FoV, we choose a thin display (Lumus OE Maximus geometric waveguide) to reduce the distance from the sensor to the eye, and a large digital sensor (Ximea CB500CG-CM) with area 36.4 mm × 27.6 mm to support a large lens array. The camera body contains over a centimeter of electronics rigidly attached behind the sensor, so we design a custom circuit board to fold the electronics off to the side. With the custom board, space for the display, and sufficient eye relief for comfort, the distance from the sensor to the eye location in our design is 35 mm.

For the lens array, we use off the shelf achromatic doublets (Edmund Optics 49-923) held together in a custom machined housing. When choosing the focal length, we are limited by sensor coverglass, which extended 1.63 mm above the sensor surface. To give sufficient room for the physical material of the apertures, we use $f = 12.5$ mm focal length lenses (Fig. 3a). With the 4.6 μm pixels on the sensor, this yields a maximum resolution under 1.5 arcmin. We choose lenses with diameter 5 mm, sufficiently above the diffraction limit. Although smaller diameter lenses could increase the f-number, resulting in longer DoF and fewer aberrations, we need a fixed area of about 1 mm between the lenses to hold them in place. As a result, smaller lenses have worse light throughput and uniformity, although this could be avoided with a monolithic lens array.

In our design, each lens only covers a small portion of the FoV, so to maximize image quality, we use Zemax to optimize the axial position and tilt of each lens in the array based on its individual FoV. The aperture location is then computed with the ray tracing approach described in Sec. 4.1, where we set the entrance pupil to be a 7 mm diameter circle 35 mm behind the sensor. Based on the optical design, we create a custom machined housing of anodized aluminum to hold the lenses in place. The housing includes physical barricades between each lens to prevent crosstalk, and ridges were cut into this material to reduce stray light and reflections within the lens tubes. A separate machined plate containing the apertures is rigidly attached, and the entire unit is mounted over the camera sensor. The resulting unit has a 45°×37° FoV, which is well matched to the FoV of the display (43.5° diagonal with 1440 × 1080 pixels per eye). Note the true FoV of our system is not rectangular due to the hex packing of circular lenses; the number reported here is the largest inscribed rectangle, which is slightly smaller than the theoretical FoV of Fig. 3b.

We create two units, one for each eye. Although we built and tested the entire binocular headset with the display, for ease of

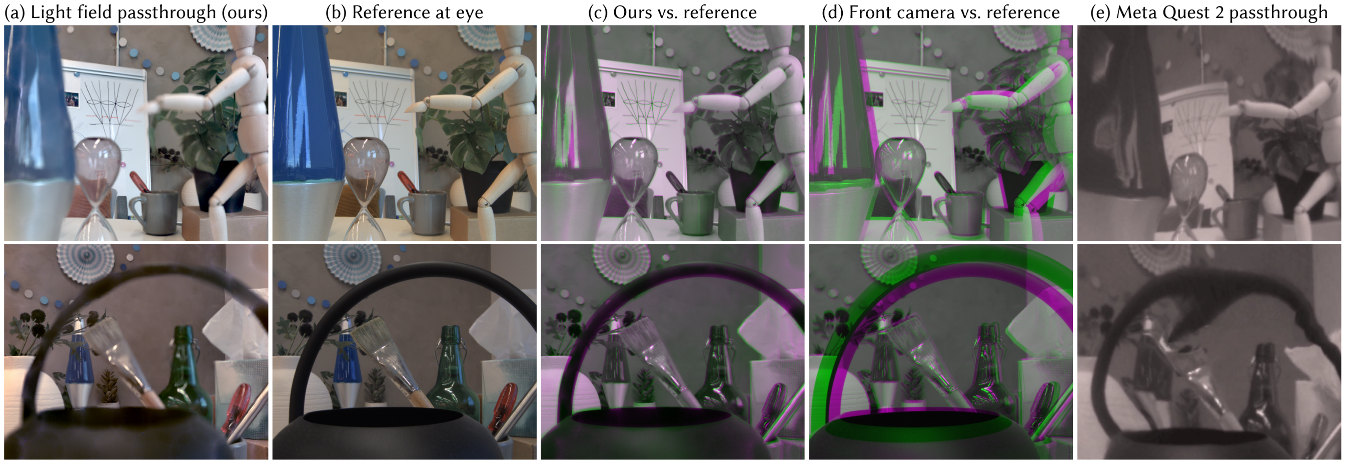

(a) Light field passthrough (ours)
(b) Reference at eye
(c) Ours vs. reference
(d) Front camera vs. reference
(e) Meta Quest 2 passthrough

Figure 6: (a) Experimental results from our light field passthrough prototype compared with (b) a reference camera placed behind the prototype at the eye location, which captures the target perspective needed for passthrough. (c) By overlaying the light field passthrough result (green) with the reference camera image (pink), we demonstrate that our camera successfully captures the correct perspective. (d) In comparison, a camera placed 35 mm **in front of the eye, the closest it can be placed given the headset form factor, captures a noticeably different perspective (green) compared to the reference camera (pink). (e) Passthrough on Meta Quest 2, a computational passthrough technique [Chaurasia et al. 2020], has significant warping artifacts.**

image capture, the results shown in this manuscript were taken with a single camera module, which is disconnected from the rest of the headset and uses the original camera housing instead of the custom board. Note that these are form factor modifications only and do not affect the image quality.

We implement all sections of the algorithm in real-time on an Nvidia 3080 GPU and achieve the following per-frame run-times for the full 7920×6004 pixel sensor:

- ~ 0.05 ms for the base algorithm (pixel re-arrangement only)
- ~ 0.6 ms for the depth estimation
- < 1 ms for the gradient domain stitching.

Additionally, although not part of the core algorithm, it takes 2.7 ms to upload each frame to the GPU and 1.4 ms for debayering. Therefore, the total time from capture to display is under 5.7 ms for a single camera. In our binocular prototype the number of cameras doubles, but we vertically bin pixels to improve sensor frame rate. Although each part of the algorithm works in real-time, for practical reasons the examples shown in this paper were processed offline in a slower Python implementation. For the depth estimation (Sec. 5.1) we use 7 planes over a 3 diopter range.

6.2 Comparisons

Since the goal of light field passthrough is to capture the perspective at the user's eye, we place a reference camera at the eye location, 35 mm behind the sensor, and compare with the reconstruction from our light field passthrough prototype (Fig. 6a,b). Although there are differences in color balance between the two cameras, by looking at occlusion boundaries, one can see that the perspective of the two cameras is close to identical, highlighted in Fig. 6c where the reference and light field passthrough results are overlayed. Note that view dependent effects like complex occluding structures and specular highlights are accurately captured by light field passthrough.

Recall that we cannot physically place a camera at the user's eye as it would be blocked by the headset. The closest we can place a traditional camera to the eye is directly in front of the headset, in this case 35 mm in front of the eye. We capture the scene from this location and overlay the image with the reference in Fig. 6d. Notice that the view from the front of the headset has significant perspective differences from the reference, particularly for close objects within arm's reach of the user. These comparisons are most easily visualized in the supplemental video.

We also provide a qualitative comparison against passthrough on the Meta Quest 2 headset (Fig. 6e) which is based on the work of Chaurasia et al. [2020]. This method has higher computational cost than our approach and creates significant distortion of the scene.

7 LIMITATIONS AND FUTURE WORK

Field-of-View. Achieving an immersive FoV requires a large lens array size (Fig. 3c), and in our implementation, we used a single large sensor behind the array. Large pieces of silicon are expensive and challenging to manufacture, limiting the practicality this approach. However, since there are gaps between the sub-aperture views in the raw sensor data (Fig. 1b), one could replace the large sensor with a collection of small sensors, essentially creating a separate camera for each lens in the array. Although this approach is more difficult to prototype, it has several advantages such as reducing cost and increasing yield by using smaller pieces of silicon, allowing for non-rectangular, non-planar arrays for more design flexibility, and reducing data bandwidth by not capturing unused pixels.

Resolution and Depth-of-Field. Despite optimizing the locations of each lens in the array, the resolution of our prototype is aberration-limited, and one can observe areas of lower resolution in the results (for example, the mannequin hand in Fig 6a). We believe the image quality could be improved with a custom optical design rather than using off-the-shelf lenses; since each lens has a narrow FoV, we expect one could achieve pixel-limited resolution using aspheres or compound optics. In addition, our prototype has limited DoF causing defocus in the foreground; this could avoided with smaller diameter lenses or autofocus, as discussed in detail in the Supplement.

Stray Light. When bright lights are present in the scene, stray light within the lens tubes can reduce contrast and cause dramatic differences in intensity between sub-aperture views. In these cases, GDIE can create additional haze in the image and may not completely remove stitching artifacts (see Supplement for an example). Improving the physical baffling of the lens array would help in these scenarios, and non-gradient domain seamless stitching techniques [Farbman et al. 2009] may remove stray light more effectively.

Eye Movement. Our design assumes that the user's eye is stationary; to account for eye movement, one could physically move the apertures in response to an eye tracker, either with physical actuation or by replacing the aperture array with a liquid crystal display controlled programatically. However, pupil movement due to eye rotation is only a few millimeters, an order of magnitude less than the visual displacement from headset thickness, so we leave this to future work.

User Study and Practicality. Our system requires additional hardware, increasing size and weight, and the image quality has artifacts compared to a traditional camera. To fully test our architecture, user studies will be necessary to determine if the benefits of accurate low-latency passthrough with correct perspective outweigh these costs. However, we note that the flat form factor of our design makes it a practical choice for HMDs, and we expect image quality to improve with future iterations of the hardware and algorithm.

8 CONCLUSION

Our optical architecture and computational pipeline exactly captures the correct perspective needed for VR passthrough. Our reconstructions are accurate with low latency, even on challenging scenes.

REFERENCES

Frank A Biocca and Jannick P Rolland. 1998. Virtual eyes can rearrange your body: Adaptation to visual displacement in see-through, head-mounted displays. *Presence* 7, 3 (1998), 262–277.

James R Bitner, Gideon Ehrlich, and Edward M Reingold. 1976. Efficient generation of the binary reflected Gray code and its applications. *Commun. ACM* 19, 9 (1976), 517–521.

Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 86–1.

Gaurav Chaurasia, Arthur Nieuwoudt, Alexandru-Eugen Ichim, Richard Szeliski, and Alexander Sorkine-Hornung. 2020. Passthrough+ real-time stereoscopic view synthesis for mobile mixed reality. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 3, 1 (2020), 1–17.

Shenchang Eric Chen and Lance Williams. 1993. View interpolation for image synthesis. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*. 279–288.

Emily K Edwards, Jannick P Rolland, and Kurtis P Keller. 1993. Video see-through design for merging of real and virtual environments. In *Proceedings of IEEE Virtual Reality Annual International Symposium*. IEEE, 223–233.

Zeev Farbman, Gil Hoffer, Yaron Lipman, Daniel Cohen-Or, and Dani Lischinski. 2009. Coordinates for instant image cloning. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–9.

Robert T. Frankot and Rama Chellappa. 1988. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10, 4 (1988), 439–451.

Henry Fuchs, Mark A Livingston, Ramesh Raskar, Kurtis Keller, Jessica R Crawford, Paul Rademacher, Samuel H Drake, Anthony A Meyer, et al. 1998. Augmented reality visualization for laparoscopic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 934–943.

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. In *Digital Optics for Immersive Displays*, Vol. 10676. SPIE, 19–35.

Todor Georgiev, Georgi Chunev, and Andrew Lumsdaine. 2011. Superresolution with the focused plenoptic camera. In *Computational Imaging IX*, Vol. 7873. SPIE, 232–244.

Phillip Guan, Eric Penner, Joel Hegland, Benjamin Letham, and Douglas Lanman. 2023. Perceptual requirements for world-locked rendering in AR and VR. *arXiv preprint arXiv:2303.15666* (2023).

Eugene Hecht. 2012. *Optics*. Pearson.

Sei-Young Kim, Joong Ho Lee, and Ji Hyung Park. 2014. The effects of visual displacement on simulator sickness in video see-through head-mounted displays. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. 79–82.

Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. 2020. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–10.

Joong Ho Lee, Sei-young Kim, Hae Cheol Yoon, Bo Kyung Huh, and Ji-Hyung Park. 2013. A preliminary investigation of human adaptations for various virtual eyes in video see-through HMDs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 309–312.

Joong Ho Lee and Ji-Hyung Park. 2020. Visuomotor adaptation to excessive visual displacement in video see-through HMDs. *Virtual Reality* 24, 2 (2020), 211–221.

Anat Levin, Assaf Zomet, Shmuel Peleg, and Yair Weiss. 2004. Seamless image stitching in the gradient domain. In *European Conference on Computer Vision*. Springer, 377–389.

Marc Levoy and Pat Hanrahan. 1996. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*. 31–42.

Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. 2000. The digital Michelangelo project: 3D scanning of large statues. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. 131–144.

Ren Jie Lin, Vin-Cent Su, Shuming Wang, Mu Ku Chen, Tsung Lin Chung, Yu Han Chen, Hsin Yu Kuo, Jia-Wern Chen, Ji Chen, Yi-Teng Huang, et al. 2019. Achromatic metalens array for full-colour light-field imaging. *Nature Nanotechnology* 14, 3 (2019), 227–231.

Andrew Maimone and Junren Wang. 2020. Holographic optics for thin and lightweight virtual reality. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 67–1.

Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. 2013. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 1–12.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Ren Ng, Marc Levoy, Mathieu Brédif, Gene Duval, Mark Horowitz, and Pat Hanrahan. 2005. *Light field photography with a hand-held plenoptic camera*. Ph. D. Dissertation. Stanford University.

Milda Park, Stefan Serefoglou, Ludger Schmidt, Klaus Radermacher, Christopher Schlick, and Holger Luczak. 2008. Hand-eye coordination using a video see-through augmented reality system. *The Ergonomics Open Journal* 1, 1 (2008).

Jannick P Rolland, Frank A Biocca, Todd Barlow, and Anantha Kancherla. 1995. Quantification of adaptation to virtual-eye location in see-thru head-mounted displays. In *Proceedings Virtual Reality Annual International Symposium'95*. IEEE, 56–66.

Thomas Schöps, Martin R Oswald, Pablo Speciale, Shuoran Yang, and Marc Pollefeys. 2017. Real-time view correction for mobile devices. *IEEE Transactions on Visualization and Computer Graphics* 23, 11 (2017), 2455–2462.

Seppe Sels, Bart Ribbens, Steve Vanlanduit, and Rudi Penne. 2019. Camera calibration using Gray code. *Sensors* 19, 2 (2019), 246.

Takeshi Shimano, Yusuke Nakamura, Kazuyuki Tajima, Mayu Sao, and Taku Hoshizawa. 2018. Lensless light-field imaging with Fresnel zone aperture: quasi-coherent coding. *Applied Optics* 57, 11 (2018), 2841–2850.

Akinari Takagi, Shoichi Yamazaki, Yoshihiro Saito, and Naosato Taniguchi. 2000. Development of a stereo video see-through HMD for AR systems. In *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*. IEEE, 68–77.

Jun Tanida, Tomoya Kumagai, Kenji Yamada, Shigehiro Miyatake, Kouichi Ishida, Takashi Morimoto, Noriyuki Kondou, Daisuke Miyazaki, and Yoshiki Ichioka. 2001. Thin observation module by bound optics (TOMBO): concept and experimental verification. *Applied Optics* 40, 11 (2001), 1806–1813.

Albert Wang, Patrick Gill, and Alyosha Molnar. 2009. Light field image sensors based on the Talbot effect. *Applied Optics* 48, 31 (2009), 5897–5905.

Lei Xiao, Salah Nouri, Joel Hegland, Alberto Garcia Garcia, and Douglas Lanman. 2022. NeuralPassthrough: Learned real-time view synthesis for VR. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817* (2018).