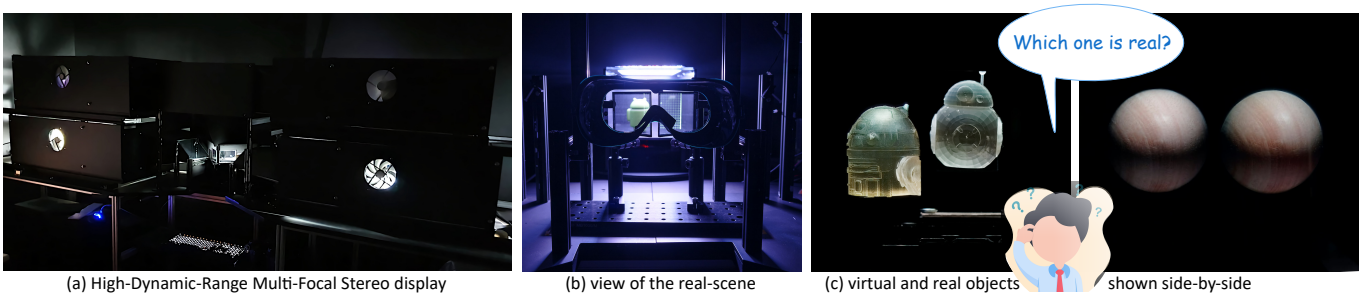
Reproducing Reality with a High-Dynamic-Range Multi-Focal Stereo Display

FANGCHENG ZHONG, University of Cambridge, United Kingdom
AKSHAY JINDAL, University of Cambridge, United Kingdom
ALI ÖZGÜR YÖNTEM, University of Cambridge, United Kingdom
PARAM HANJI, University of Cambridge, United Kingdom
SIMON J. WATT, Bangor University, United Kingdom
RAFAŁ K. MANTIUK, University of Cambridge, United Kingdom



(a) High-Dynamic-Range Multi-Focal Stereo display (b) view of the real-scene (c) virtual and real objects shown side-by-side

Fig. 1. We built a High-Dynamic-Range Multi-Focal Stereo display (a) which allows for a direct comparison with a physical scene located in front of the observer (b). The display can reproduce real-world 3D objects with accurate color, contrast, disparity, and a range of focal depth, making it hard to distinguish between real and virtual scenes (c).

With well-established methods for producing photo-realistic results, the next big challenge of graphics and display technologies is to achieve perceptual realism — producing imagery indistinguishable from real-world 3D scenes. To deliver all necessary visual cues for perceptual realism, we built a High-Dynamic-Range Multi-Focal Stereo Display that achieves high resolution, accurate color, a wide dynamic range, and most depth cues, including binocular presentation and a range of focal depth. The display and associated imaging system have been designed to capture and reproduce a small near-eye three-dimensional object and to allow for a direct comparison between virtual and real scenes. To assess our reproduction of realism and demonstrate the capability of the display and imaging system, we conducted an experiment in which the participants were asked to discriminate between a virtual object and its physical counterpart. Our results indicate that the participants can only detect the discrepancy with a probability of 0.44. With such a level of perceptual realism, our display apparatus can facilitate a range of visual experiments that require the highest fidelity of reproduction while allowing for the full control of the displayed stimuli.

CCS Concepts: • **Computing methodologies** → **Perception**.

Additional Key Words and Phrases: computational displays, perception, visual experiments, high dynamic range, perceptual match, quality assessment

ACM Reference Format:
Fangcheng Zhong, Akshay Jindal, Ali Özgür Yöntem, Param Hanji, Simon J. Watt, and Rafał K. Mantiuk. 2021. Reproducing Reality with a High-Dynamic-Range Multi-Focal Stereo Display. *ACM Trans. Graph.* 40, 6, Article 241 (December 2021), 14 pages. https://doi.org/10.1145/3478513.3480513

1 INTRODUCTION

Photorealism in computer graphics — rendering images and animations that appear the same as photographs or cinematographic movies — has matured to the point that it is now widely used in industry. Yet, this approach places an upper limit on the realism achieved by a photograph. Emerging display technologies can deliver high dynamic range (HDR), accurate color reproduction, and a close approximation to the full set of real-world cues of 3D scenes (including focal depth cues). Together, such display technologies can potentially exceed the realism of photographs and bring us closer to what we define as perceptual realism — displaying content that is perceptually indistinguishable from real-world 3D scenes.

Perceptual realism puts very strict requirements on the quality of reproduction. To make the task feasible, we aim for a visual reproduction of a static scene encompassing a moderate field of view (27°×21.8°) and seen from a fixed viewing position (no motion parallax). Such a scene can in principle be reproduced with

Authors' addresses: Fangcheng Zhong, Dept. of Computer Science and Technology, University of Cambridge, United Kingdom, fangcheng.zhong@cst.cam.ac.uk; Akshay Jindal, Dept. of Computer Science and Technology, University of Cambridge, United Kingdom, aj577@cst.cam.ac.uk; Ali Özgür Yöntem, Dept. of Computer Science and Technology, University of Cambridge, United Kingdom, aoy20@cam.ac.uk; Param Hanji, Dept. of Computer Science and Technology, University of Cambridge, United Kingdom, pmh64@cam.ac.uk; Simon J. Watt, School of Psychology, Bangor University, United Kingdom, s.watt@bangor.ac.uk; Rafał K. Mantiuk, Dept. of Computer Science and Technology, University of Cambridge, United Kingdom, rafal.mantiuk@cl.cam.ac.uk.

0730-0301/2021/12-ART241
https://doi.org/10.1145/3478513.3480513

perceptually-realistic fidelity if we can achieve sufficient quality in terms of spatial resolution, color gamut, dynamic range, and depth cues. The fact that human perception integrates across different cues, creating a 'holistic' percept, raises the possibility that almost inevitable small differences to the real world in terms of individual attributes may not be noticeable provided the other display capabilities are of sufficient quality.

The first objective of our work is to build a display apparatus and a 3D scene acquisition and rendering system that combines high spatial resolution with accurate colors, luminance levels, and cues to 3D structure (including focal distance). Our display apparatus combines four custom-built HDR displays into a single-viewer two-focal plane stereoscopic display. It can deliver a brightness level up to 3000 cd/m^2 and below 0.01 cd/m^2, a spatial resolution of at least 85 pixels per degree at a viewing distance of 462 mm, a color gamut of BT.709, correct disparity, and variations in focal depth from 462 mm (2.16 D/diopters) to 740 mm (1.35 D). These capabilities are sufficient to reproduce a small scene inside a box of size 200 mm × 160 mm × 300 mm (width × height × depth) with levels of realism that exceed what existing display technologies can offer. Furthermore, the display is constructed in such a way that a viewer can simultaneously, or selectively, see a physical box containing real objects and compare them with displayed ones in the same spatial position. This enables a set of new perceptual experiments that have not been possible before. To deliver high-quality content for such a display, we create a system for acquiring, reconstructing, and rendering 3D scenes with a lumigraph [Gortler et al. 1996] (light field with a proxy mesh). The system involves the capture of multi-exposure image stacks from multiple viewpoints with a high-resolution DSLR camera, camera pose estimation with photogrammetry, color calibration with a spectrometer, proxy mesh registration with differentiable rasterization, lumigraph view synthesis with view-dependent UV maps, multi-focal rendering with linear depth filtering, and a custom-designed focal plane calibration to compensate for different viewing positions of observers.

Our second objective is to use this system to visually reproduce a small stationary object at a close distance to the observer (0.5 m) with a high fidelity such that it can be confused with a physical 3D object. The fidelity of the reproduction is confirmed by a visual Turing test [Banks et al. 2016] with a strict criterion: the virtual scene must not be visually different in any respect from the real scene. This is tested in a three-interval-forced-choice (3IFC) experiment in which we ask naive observers to choose a scene that appears different when presented with two real and one virtual scenes, or one real and two virtual scenes. This way we evaluate realism objectively and eliminate subjective interpretations. The attempt at this challenge provides us insights to better understand the conditions necessary to achieve perceptual realism. In the long term, we foresee this approach as an important step in the study of future display technologies, including AR and VR, to determine what display capabilities are most critical in achieving perceptual realism. Our display apparatus can also be useful in studies of material perception, color appearance, and depth perception, in which realistic objects and scenes need to be faithfully reproduced.

We make the following contributions: (i) We built a novel display apparatus with an imaging system that is capable of capturing and reproducing all essential perceptual cues of a static scene of moderate size and with the capability to switch between viewing real and displayed scenes with the observer in the same position. (ii) To our knowledge, our work is the first that achieves a close perceptual match between a real-world 3D object and its displayed counterpart in both geometry and appearance. We experimentally demonstrate that our display apparatus can pass a visual Turing test — naive observers can only distinguish between real and displayed 3D objects with a probability of 0.44. In contrast to previous work [Borg et al. 2012; Masaoka et al. 2013; Meyer 1986], we achieved this with binocular viewing of a near object, and without any optical degradation of the real scene.

2 RELATED WORK

Obtaining realistic results has been one of the main pursuits of computer graphics and in particular of rendering. Global illumination and physically based rendering allowed for accurate simulation of light [Goral et al. 1984]. When combined with tone mapping methods [Eilertsen et al. 2017], such as simulation of lens glare [Ritschel et al. 2009] and camera response [Reinhard et al. 2002], these techniques can produce photorealistic images, indistinguishable from photographs of real-world scenes. However, since the focus of our work lies beyond photorealism, we review the studies that attempted to achieve perceptual realism by matching a virtual scene with a physical one.

Meyer [1986] was the first to compare rendering shown on a display with a real scene in an experiment. The participants saw the real scene and a CRT screen with its reproduction side by side, via viewfinders of two cameras with telephoto lenses. Additional Fresnel lenses were added to enlarge the viewfinder images so that they could be seen from 112 cm. Despite the lack of binocular depth cues and the low resolution of the CRT screen, the authors reported that neither naive observers nor experts could tell which image was computer generated. Although this was an impressive result, it was helped by the degradation of the real-scene images, due to lens distortions, and their small size (9.2×9.2 cm seen from 112 cm, or 4.7°).

Borg et al. [2012] reported a graphics Turing test experiment, in which they successfully reproduced the result of Meyer without the need to see the stimuli via a viewfinder. The participants viewed either a real object (a pyramid or a sphere), or a display seen through a small aperture in a 2 m long box. The stimuli were viewed with one eye. Also, because the authors could not achieve the required dynamic range on their display they asked the participants to view the images from 10 cm away from the box in a non-dark room (50 lux) so that the display black level was masked by glare in the eye and adaptation.

Masaoka et al. [2013] measured how the impression of realism is degraded with the reduction of resolution. The authors conducted a pairwise comparison experiment, in which one of the conditions was a real scene and the other conditions were images of gradually reduced resolution. The results of comparisons were scaled using a Bradley-Terry model to give a measure of the sense of realness, proportional to just-noticeable-difference (JND) units. The images and the real scene were seen through a synopter so there were no

binocular disparities, and the distance was 480 cm to ensure sufficient angular resolution and minimise the influence of variations in focal distance. The study found that a resolution between 60 and 120 cycles per degree is required to achieve the perceived realism of a real scene.

None of the above studies attempted to reproduce binocular depth cues but instead reduced their influence by using large viewing distances and optics. These studies also reported difficulties in reproducing the real-world dynamic range. Both of these aspects were addressed in the study of Vangorp et al. [Vangorp et al. 2014], in which the virtual scene was reproduced on an HDR display (SIM2 HDR47E) seen through a stereoscope, albeit at low resolution (30 ppd). The goal of the study was to measure how binocular disparity and contrast contribute to realism, in a manner similar to Masaoka et al.'s study of resolution. The task was to compare two displayed scenes, each with a certain amount of both contrast and disparity modification, and choose the one closer to the real scene. The participants could look at the real scene at their discretion, but it was not included in the compared conditions, so the experiment could not test for a perceptual match. The authors found that the participants were more sensitive to changes in contrast than in disparity, and selected as more realistic either natural or moderately enhanced contrast.

In addition to the above-mentioned visual Turing-test experiments, a comparison with a real scene has also been used to evaluate reproduction of brightness [McNamara 2005] and tone mapping [Yoshida et al. 2006], but these studies did not attempt to achieve a perceptual match with a real scene.

Although the studies of Meyer, Borg et al., and Masaoka et al. reported a perceptual match of the display and real scenes, they were achieved only in monocular view or using optics that degraded visual quality of the real scene. Our work aims to go beyond these efforts. We reproduce all visual cues, including depth and dynamic range, and match a real object seen at a small viewing distance, and with no optical aberrations.

3 HDR-MF-S DISPLAY

The main objective of the design of our HDR-MF-S display is to maximize the visual quality and realism of the displayed images for all the following capability dimensions: physical luminance, dynamic range (contrast), color gamut, binocular and focal depth cues. The goal is to deliver all these capabilities altogether with sufficient qualities rather than focusing on maximizing a single one. While there are several fundamentally distinct approaches to 3D display architectures, not all of them meet the requirements for our objective. For example, accurate depth cues, matching light distributions in the real world, can be potentially achieved with holographic [Lucente 2012] or light field [Surman and Sexton 2012] displays. However, the current state-of-the-art of these technologies does not allow us to achieve the field of view, color accuracy, resolution, or dynamic range required for perceptual realism. Reproducing a 4-dimensional light field of sufficient size and quality with these technologies requires control over billions of pixels, which is currently infeasible. However, if we can either stabilize or track the viewing position, the subspace of a light field that we need to reproduce is much smaller, making it possible to build a display of required capabilities.

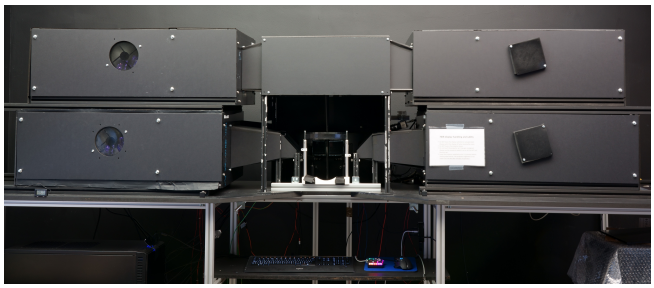


Fig. 2. The front view of the display.

One approach to producing the required light distribution, given either fixed or known eye position, is to use a stereoscopic multi-focal display [Akeley et al. 2004]. In such displays the eye sees the sum of light from multiple superimposed planes at different focal distances. Such displays can effectively drive accommodation to any point between the planes if the plane separation is small enough (~0.6 D to ~0.9 D) [MacKenzie et al. 2012, 2010], while retaining desirable capabilities of conventional displays (resolution, color gamut). Moreover, this uncomplicated design, without any refractive or diffractive optical components in the viewing path, generates images without additional optical distortions. This is in contrast with vari-focal displays [Dunn et al. 2017] or near-eye light field displays [Huang et al. 2015], which are likely to introduce noticeable aberrations. One important limitation of a multi-focal display is that the addition of focal planes reduces dynamic range. The additive nature of the beam-splitters elevates black level, and their transmission limits the peak brightness of each plane. We address this problem by combining a multi-focal stereoscopic display design with high-dynamic-range displays, making a high-dynamic-range multi-focal stereoscopic (HDR-MF-S) display. In the following subsections, we explain the details of the design of our HDR-MF-S display and how it achieves the capability dimensions that we desire. We also provide additional descriptions of each component of our display setup with CAD drawings in the supplementary materials.

3.1 Apparatus overview

Figure 2 shows a photograph of the front view of our display apparatus. The apparatus comprises three main components as shown in Figure 3: a Wheatstone stereoscope with four high-dynamic-range displays and two focal planes; a real-scene box in front of the observer that is seen through a pair of beam-splitters; and a motorized camera slider capable of capturing dense horizontal light fields of the real-scene box. In this setup a small physical scene is arranged in the real-scene box. This box normally faces the observer, but can be rotated to face the camera rig in order to capture its light field as shown in Figure 3. When facing the observer, the real scene and its rendered counterpart are spatially superimposed. We can instantly switch between the real and displayed scenes by controlling the lights in the real-scene box and the display. We discuss the details of each component in the following subsections.

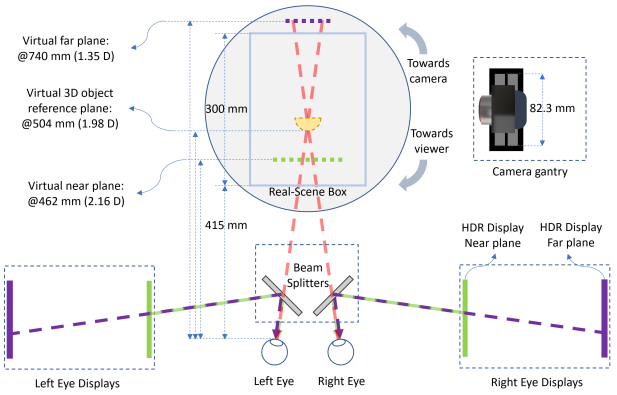


Fig. 3. Schematic of the high-dynamic-range multi-focal stereo display apparatus. (Note that to simplify the schematic, not all the folding mirrors and beam splitters are shown.) The apparatus creates two image planes (green and blue dashed lines inside the real-scene box) per eye and the observer sees respective images via beam splitters. The real-scene box is observed through the same beam splitters. The real-scene box is on a manually rotating platform moving toward a fixed capturing position or a fixed display position. The camera gantry is on another manually movable platform (not shown in the figure) which can move towards or away from the real-scene box allowing coarse adjustment of the field of view.

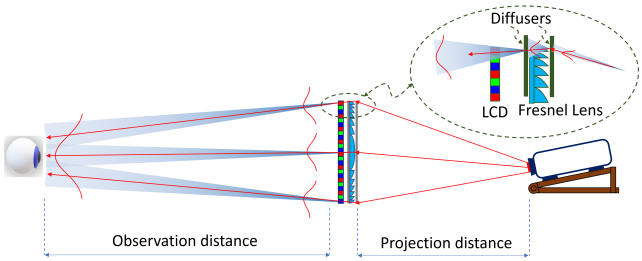


Fig. 4. Each HDR display comprises a projector acting as a backlight for an LCD panel with factory backlighting removed. A Fresnel lens sandwiched between two narrow-angle diffusers, with scattering angles of 10 and 5 degrees. An image from the projector is formed on the first diffuser acting as the backlighting of the LCD. The Fresnel lens helps to steer the backlighting toward the eye uniformly. The second diffuser prevents reflections between LCD glass and Fresnel lens substrate.

3.2 HDR displays

The key feature of our display is the capability of reproducing a high dynamic range, with a peak luminance of 3000 cd/m^2 and the black level much below 0.01 cd/m^2. Such a low black level practically eliminates any stray light in areas of an image that should remain black. The HDR reproduction is delivered by four projector-based dual-modulation displays, similar in design to those used in one of the first HDR displays [Seetzen et al. 2004] but with multiple improvements, explained below.

Each such HDR display consists of an IPS LCD panel (9.7" LP097QX1 2048×1536) from iPad3 with the backlight removed, and substituted

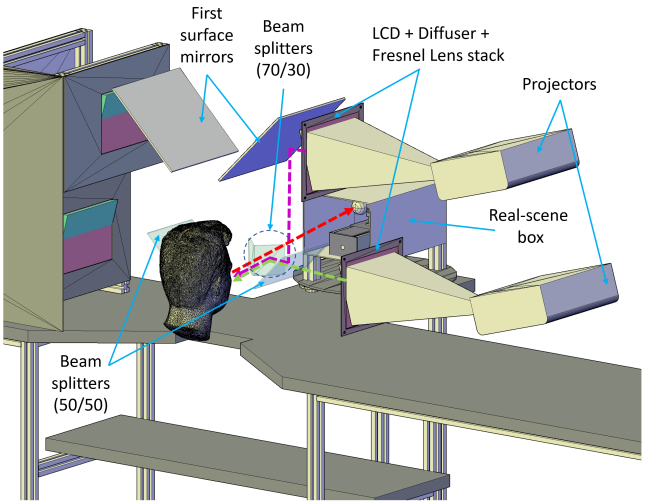


Fig. 5. The schematic showing the light paths from two display focal planes (green and purple dashed lines) and from the real-scene box, for the right eye. The red dashed line shows the viewing direction of the observer towards the real-scene box. The line colors are consistent with Figure 3.

by a DLP projector (Acer P1276) with its color wheel removed. For realizing the image from the projector, we used a Fresnel lens (Comar Optics) sandwiched between two narrow-angle diffusers (Luminit), with scattering angles of 10 and 5 degrees. This prevented double reflections between the LCD glass and the Fresnel lens substrate. The selection of scattering angles gave the best trade-off between the light efficiency and the uniformity of the display. Figure 4 depicts the optical structure for a single projector setup. The projector was positioned on a tilted ramp to reduce the keystoning effect. To further maximize the light efficiency, the focal lengths of the near- and far-plane Fresnel lenses were selected as 254 mm and 279 mm, respectively, to focus the light from each plane towards the eye of the observer.

The software for controlling each display implemented the standard two-spatial-modulator factorization algorithm [Seetzen et al. 2004] running on a GPU. However, we took special care to achieve accurate geometric alignment and high color accuracy. The geometric alignment was achieved by taking images with a DSLR of a calibration pattern (a grid of points) displayed separately on the LCD and the DLP and then aligning them using homography and mesh-based warping. The point-spread-function of the DLP was measured for the same grid of points and approximated with a Gaussian function. The colorimetric calibration was achieved by measuring the color ramps with a spectro-radiometer (Specbos 1211) and fitting a gamma-offset-gain model to the LCD panel and using a dense look-up table for the DLP. The dense look-up table was necessary as the response of the projector was non-monotonic after removing the color wheel. The effective bit-depth of both displays was increased to 10 bits by bit-stealing (DLP) and spatio-temporal dithering (both DLP and LCD). The uniformity of the display was compensated by taking an image with a DSLR and using it for compensation of the DLP image.

Fig. 6. Frames of the glasses with an IR LED. The participants were asked to wear these frames to track their head position.

3.3 Focal planes and optics

To vary focal distance, similar to a multi-focal display [Akeley et al. 2004], our display can generate images at two focal planes, at the distances of 462 mm (2.16 D) and 740 mm (1.35 D) from the viewer, providing a 0.81 diopter separation between the planes. The separation was selected to ensure that the images shown on two planes provide cues for accommodation for any distance between the two planes [MacKenzie et al. 2010]. Such distances also ensure a resolution of at least 85 pixels per degree for the observer. These distances are adjustable by moving the HDR displays on their mounting rails.

Figure 5 shows the optical paths for near and far virtual images on the right-hand side. The image of the far plane is formed by reflecting the real image of the top right HDR display through a mirror, and two beam-splitters. The purple dashed line in the figure indicates its optical path. The near-plane image is formed by reflecting the real image of the bottom HDR display from a single beam-splitter, depicted by the green dashed lines. This is symmetrical for the left-hand side of the setup. We opted for this simple optical design without any refractive [MacKenzie et al. 2010] or varifocal [Chang et al. 2018] optics to avoid aberrations, which would introduce detectable imperfections and also reduce the dynamic range due to scattering of the light. The real-scene box is observed through 70R/30T (reflection/transmittance, Edmund Optics, 64-409) beam-splitters, located in front of the observer's eyes. The red dashed line shows the viewing direction through these beam-splitters. This reflection/transmittance ratio was selected to achieve a higher brightness of the display. The second beam-splitter 50R/50T (weidner-glas.de) on the side is used to combine the images from far and near planes. Since the system has several optical paths crossing each other, we had to enclose all the image delivering paths separately to avoid cross-talk images. At the optical exit where the observer views the scene and the displays, we placed a chin-rest and forehead-rest to fix the viewing direction and limit head movements. We also placed blinders on either side of the chinrest posts to prevent direct line of sight of the near plane LCD screens.

Multi-focal plane displays are very sensitive to misalignment due to head-movement and often require either bite-bars [MacKenzie et al. 2010] to eliminate such movements or active correction in rendering through eye tracking [Mercier et al. 2017]. We aimed to build a setup similar to the latter using an IR LED fixed onto a glasses frame without lenses (Figure 6). The observers were asked to wear the frame while viewing, and the LED was tracked using a high frame rate machine vision camera (iDS UI-3140CP), with 25 mm C-mount lens (Fujinon HF25HA-1B) and a visible light filter. This allowed us to track the observer's head position in real-time. We later use the data from the head tracker in our experiment (Section 6) to determine the invalid trials.

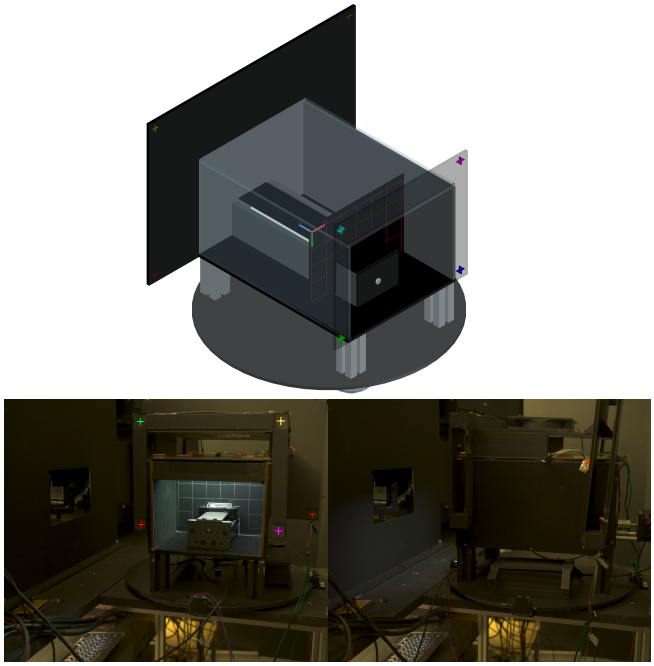


Fig. 7. The front and side view of the real-scene box and schematic of the calibration target inside. The calibration target has a grid of four-by-six squares of the size 30 mm x 30 mm, which defines a world coordinate system. The red, green, and blue arrows in the figure represent the origin and orientations of the X, Y, and Z axes, respectively. We define the upper-left corner of the grid as the origin for the X and Y axes and the target placed at the front location as the $Z = 0$ plane.

3.4 Real-scene box

The real-scene box has the inner dimensions of 200 mm × 160 mm × 300 mm (width × height × depth), matching the physical size of the LCD screens and accommodating the optical separation distance between the far and near plane HDR display apertures. It was made of black acrylic, which was covered on the inside with high absorption blackout material (Thorlabs: Black Flocked Self-Adhesive Paper). The ceiling was fitted with an LED array light source with 225 individually addressable RGB LEDs (WS2812B). The real-scene box was fixed on a platform, supported by ball transfer units, allowing it to be freely rotated towards the observer for viewing, or towards the camera for light-field capture, as shown in Figure 7. The real-scene box rotation was fixed in either of the two positions using custom magnetic mounts.

To facilitate several calibration procedures for our imaging system (Section 4), we defined a world space coordinates for the real-scene box. We placed a removable calibration target on a gantry plate inside the real-scene box, as shown in Figure 7. The gantry (Oozenest, 250 mm C-Beam Linear Actuator) can be controlled to move the target freely from the entry of the real-scene box to its end. The calibration target had a grid of four-by-six squares of the size 30 mm x 30 mm. We used the grid to define a world coordinate space, as shown in Figure 7.

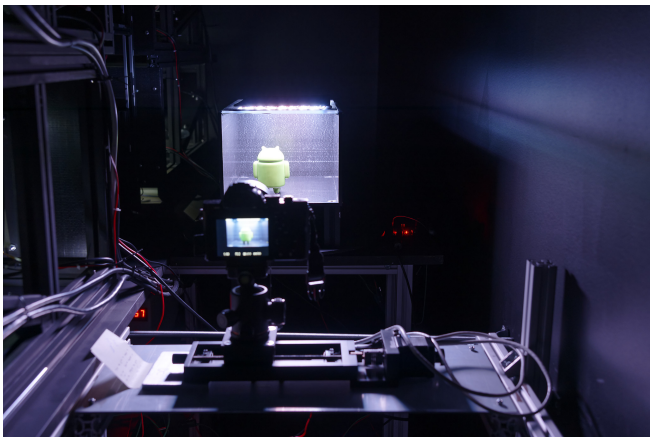


Fig. 8. The data camera and motorized slider for light field capture.

In addition to the calibration target, the real-scene box also included eight cross-shaped calibration markers placed outside the box, as shown in Figure 7. The markers were used as a reference points to register the camera pose when the calibration target inside the box had to be removed. The markers were carved on the two foamboards and illuminated by an RGB LED (WS2812B) with a diffuser to improve their visibility.

3.5 Data camera for light field capture

To capture a horizontal light field of real-scene box, we mounted a Sony α7R3 mirror-less camera with a Sony G OSS zoom lens (focal length 24-105 mm) on a motorized camera slider (Figure 8) at a distance of 415 mm from the real-scene box, similar to the distance from the viewing position to the real-scene box. The camera slider traversed a baseline of 82.3 mm with an accuracy of 5 µm.

4 HDR-MF-S IMAGING & RENDERING SYSTEM

To achieve perceptual realism, we need not only a display capable of reproducing all relevant cues, but also an imaging and rendering system, which can capture a real scene and display it with sufficient quality. Most importantly, the rendered scene should match the viewpoint of the observer. Our system is currently limited to processing scenes of relatively simple or known geometry, but can handle complex non-Lambertian materials and high-dynamic-range illumination.

Figure 9 shows a diagram of the HDR-MF-S imaging and rendering system. We start with the capture of a horizontal HDR light field, which is color-calibrated for the spectra of the scene illumination (Section 4.1). Next, we employ photogrammetry to perform a 3D reconstruction and estimate camera matrices (Section 4.2). After that, we apply a differentiable rasterizer to register a proxy mesh of the main object with its silhouette in each HDR light field image (Section 4.2), so we can project the fitted mesh to each light field image to obtain a view-dependent UV map and texture. Before rendering, we find the position of each focal plane of the display with respect to the eye position and the calibration target in the real-scene box (Section 4.3). Finally, we integrate lumigraph view synthesis with linear depth filtering [Akeley et al. 2004] to render the final scene on our HDR-MF-S display (Section 4.4).

We found lumigraph to be the most suitable 3D representation for our purpose as it models non-Lambertian surfaces, is robust to processing high-resolution textures, and performs rendering in real time. We have also experimented with dense light fields, either captured or reconstructed using neural radiance fields [Mildenhall et al. 2020], but they did not match the quality required for perceptual realism, which poses a lower tolerance for artifacts (such as blur, noise, and distortion) and a higher demand for capacity to process high-resolution (8k) images. Instead, we combine photogrammetry and differentiable rendering to align known geometry with the captured HDR images to reconstruct a lumigraph.

4.1 HDR light field capture

Using our data camera discussed in Section 3.5, we first capture a high-resolution (7360 x 4912 pixels) light field consisting of 16 views with a separation of 5 mm between them. For each camera view, we capture an HDR exposure stack consisting of up to five RAW images spaced two stops apart in exposure time and ISO of 100. We merge the RAW images to increase the dynamic range and reduce noise using a Poisson photon noise estimator [Hanji et al. 2020]. Next, we demosaic the merged images using the *DDFAPD algorithm* [Menon et al. 2006]. To calibrate for colors, we measure the spectra of a color checker passport (X-Rite) positioned inside the real-scene box with a spectroradiometer (Specbos 1211, Jeti). Then, we compensate for the measured spectral transmission of the 70/30 beam-splitter and recover trichromatic coordinates using the CIE XYZ 1931 colour matching functions. The XYZ color coordinates are used to find the matrix that transforms from native camera linear RGB space into CIE XYZ and which results in the smallest RMSE of DeltaE 2000 color differences. The white patch in the color checker is used for white balance. Finally, we apply the matrix to convert the merged HDR images from their native camera linear RGB space to the BT.709 space used by our display.

4.2 Lumigraph reconstruction

The objective of this stage is to construct a lumigraph [Gortler et al. 1996] (a light field projected on a proxy geometry), represented by a proxy mesh and view-dependent UV maps and textures, of the captured scene.

Photogrammetry. We first use *Meshroom* [AliceVision 2018], a photogrammetry software, to perform a multi-view stereo reconstruction of the scene. We supply Meshroom with the HDR light field images captured from the gantry and additional single-exposure images captured with the camera mounted on a Magic Arm (Manfrotto) and positioned at multiple locations around the front of the real-scene box. These additional images are necessary for the 3D reconstruction but are not used for textures. After the reconstruction, Meshroom returns a noisy scene mesh (including the main object, the real-scene box, and the calibration markers, etc.) with estimated camera extrinsic and intrinsic matrices. Note that at this stage, the scene mesh is in an arbitrary local camera space. The camera matrices are also calculated with respect to this space. We

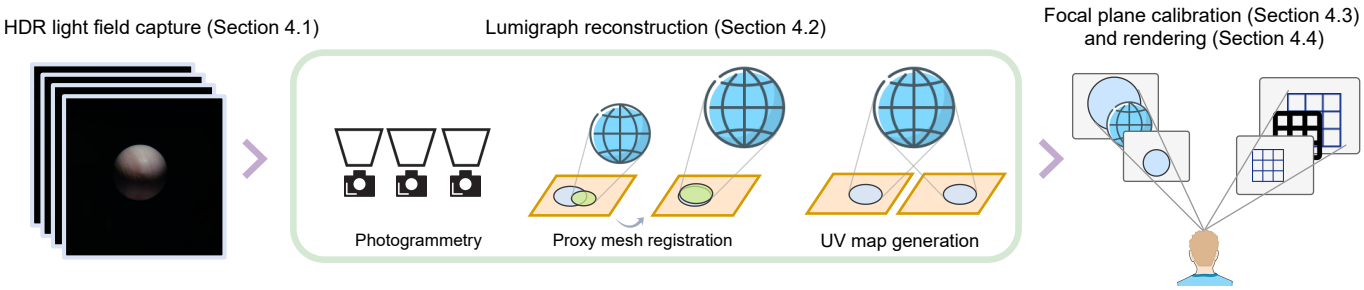


Fig. 9. The process of capturing and rendering contents for our HDR-MF-S display. Refer to Section 4 for the explanation.

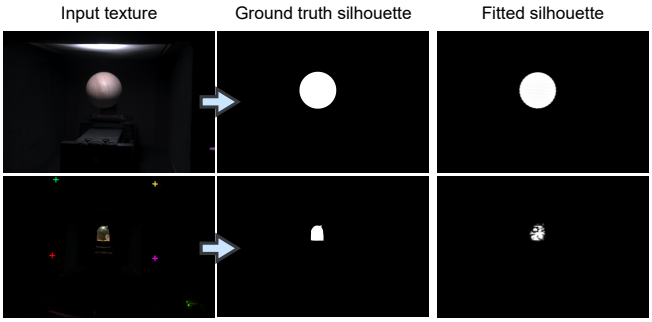


Fig. 10. Results of the fitted silhouettes of the proxy mesh after registration optimized by differentiable rasterization.

record the coordinates of each reconstructed calibration marker in local space, which we later use for a coordinate transform.

Proxy mesh registration and UV map generation. The mesh reconstructed from photogrammetry does not meet the accuracy of perceptual match required by our experiment. Hence, we choose to experiment with objects with simple or known geometry and pre-generate the mesh files, as mesh reconstruction is not the main focus of this work. However, we still need to register the mesh to the correct coordinates. It is crucial to ensure that the projected silhouette of the registered mesh is near-identical to the ground truth. Otherwise, the rendering would appear distorted once we project the mesh onto light field images to construct the lumigraph. We employ *SoftRas* [Liu et al. 2019; Ravi et al. 2020], a differentiable rasterizer, to find an optimal spatial transformation to align the mesh with the silhouettes in captured images. Specifically, the optimal parameters of a spatial transformation $\mathbf{T}$ including scaling, rotation, and translation can be found by

$$\arg\min_{\mathbf{T}} \sum_{i} ||\mathbf{R}(\mathbf{T}(M), C_i) - I_i||, \tag{1}$$

where $\mathbf{R}$ is a differentiable renderer that rasterizes a grey-scale silhouette image, M is the unregistered mesh, C_i is the i-th camera matrix, and I_i is the extracted ground-truth silhouette from the i-th camera view. We apply the *GrabCut algorithm* [Rother et al. 2004] to extract the ground-truth silhouettes of the main object. Figure 10 shows the results of the silhouette fitting. After the registration of the proxy mesh, we generate the UV coordinates by projecting the mesh vertices onto each HDR texture using the camera matrices obtained from photogrammetry.

Local-to-world coordinate transformation. To facilitate the following calibration steps, it is convenient to have the scene geometry represented in world coordinates expressed in physical units (meters). To do this, we determine the coordinates of the calibration markers in both local space (Section 4.2) and world space (Section 3.4) and apply the *orthogonal Procrustes algorithm* to find an optimal change-of-coordinates transformation from the local to the world space.

4.3 View-dependent focal plane calibration

Both pairs of display focal planes must be well-aligned with the positions of the observer's eyes to correctly align the two focal planes and match the scene shown in the real-scene box. To map the coordinates of each display to the world coordinates of the real-scene box, we perform a manual focal plane calibration. As different observers have different inter-pupillary distances (IPDs) and may put their heads at different positions, this calibration needs to be performed per observer.

During the calibration, the observer is asked to put their head on the chin rest and press against a rigid forehead rest. The forehead rest provides additional stability and limits head movements. As shown in Figure 11, each eye is presented with four crosses on one of the HDR displays. They move the four crosses to align them to the corresponding specified crossings of the calibration target in the real-scene box. The observers perform this alignment for each of the two focal planes per eye and for the calibration-target positioned at two different depths. The gantry inside the box moves the target to their desired locations. After this calibration, we obtain a correspondence of eight points in world space and in image space. They are used to find the transformation from the 2D coordinates on each focal plane (an image shown on each HDR display) to the world coordinates. We use the *direct linear transformation algorithm* (DLT) [Sutherland 1974] to find a rendering matrix M which maps the world coordinates to the clip space for each focal plane. Finally, we apply an RQ decomposition to decompose the rendering matrix into a view (extrinsic) matrix V and a projection (intrinsic) matrix P, i.e. $M = PV$. With the view matrix we are able to compute the observer's view (eye) positions and orientations, which is required for the lumigraph view synthesis and multi-focal decomposition in the rendering stage.

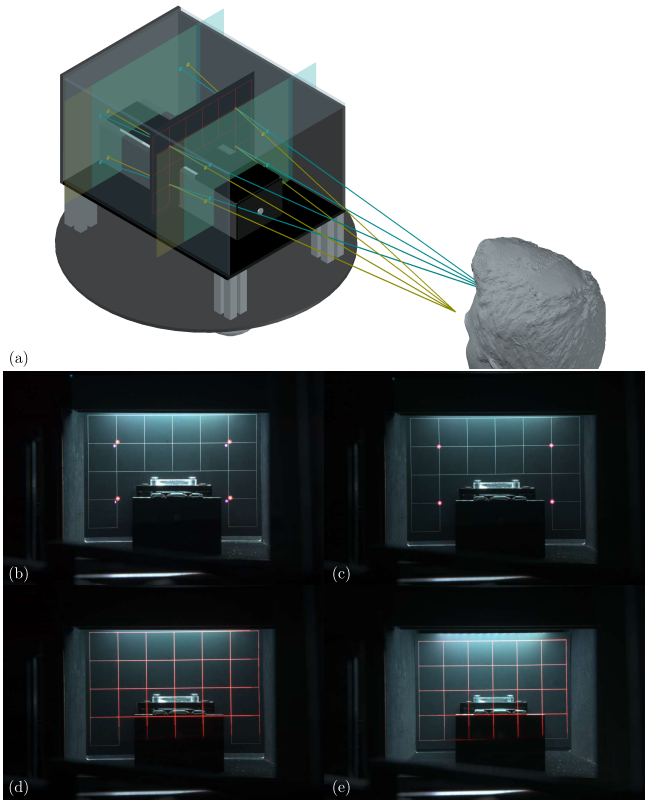


Fig. 11. (a) Schematic of the focal plane calibration. We use yellow and cyan to indicate the view of the left and right eye. (b, c) Left-eye view of the focal plane calibration interface. Observers drag the red (near plane) and pink dots (far plane) to align with the corresponding positions on the calibration target. (d, e) Rendering of the calibration grids at different gantry positions after calibration.

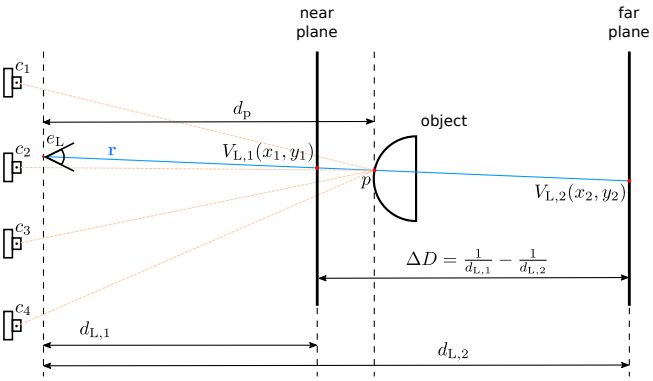


Fig. 12. The radiance computation for the near and far focal planes for the left eye. $c_1, \ldots, c_4$ are the positions of data cameras. e_{L} is the viewing position of the left eye.

4.4 Multi-focal lumigraph rendering

To find the value of each pixel of the near and far display focal planes, we use lumigraph rendering [Gortler et al. 1996], combined with linear depth filtering in the diopter space [Akeley et al. 2004]. We choose simple linear filtering as our test scene does not contain any occlusions, which would require more advanced methods [Mercier et al. 2017; Narain et al. 2015; Yu et al. 2019]. Specifically, the value of the pixel (x, y) on the j-th focal plane (1 – near, 2 – far) for the left eye (index L) is computed by filtering across the focal planes and cameras (similarly for the right eye):

$$V_{L,j}(x,y) = \underbrace{\frac{|D_{\mathrm{p}} - D_{\mathrm{L,j}}|}{\Delta D}}_{\text{linear depth filtering}} \underbrace{\sum_{k=1}^{K} T_k(u_k, v_k)\, w_k}_{\text{view synthesis across } K \text{ camera views}}, \tag{2}$$

where the symbols are illustrated in Figure 12. We use lower case symbol d to represent distances in meters and upper case symbol D represent distances in diopters, so that $D = 1/d$. In particular, $D_{\mathrm{L,j}}$ is the diopter of the j-th focal plane from the viewing position e_L. D_{p} is the distance (in diopters) of the intersection point p of the ray $\mathbf{r}$ with the object, where $\mathbf{r}$ originates from e_L and passes through pixel (x, y). ΔD indicates the diopter difference between the near and far focal planes. $T_k(u_k, v_k)$ represents the value of the HDR texture associated with the data camera k for the texture coordintes (u_k, v_k) at the intersection point p. We calculate T_k by rasterizing the texture-mapped registered mesh (Section 4.2) with the rendering matrices generated during the focal plane calibration (Section 4.3). The texture is filtered with standard mipmap. The value of w_k is the weight associated with each data camera. As we assume a static eye position, we always select the nearest neighbor in our current implementation to avoid blur artifacts:

$$w_k = \begin{cases} 1, & \text{if } ||e_L - c_k|| = \min_j ||e_L - c_j||, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where the values of e and c (data camera origins) are obtained from the focal plane calibration (Section 4.3) and lumigraph reconstruction (Section 4.2) respectively.

5 RESULTS

Although it is difficult to convey the three-dimensionality and color appearance of the scenes shown on our display using photographs, in this section, we include a few to demonstrate some of its characteristic capabilities. We captured images of several displayed and real objects using a Sony α7R3 camera with a 55 mm lens (SEL55F18Z). We set the aperture to F9.5 so that its diameter matched the expected pupil diameter for our scene (5.8 mm). We also performed the focal plane calibration (Section 4.3) for the viewing position of the camera.

Figure 13 demonstrates a close perceptual match between the real and virtual objects achieved by our system. The accurate spatial alignment of the virtual object overlaying the physical object demonstrates the perceptual match in geometry (Figure 13(a)). We also achieved a close match in appearance and shading (see the overlapping shadows and specular reflections in Figure 13(a) and the side-by-side comparison in Figure 13(c)). With such a level of precision, we are able to show many mixed-reality effects that would not be possible otherwise such as changing the hue of the physical object without changing the shadows or textures.

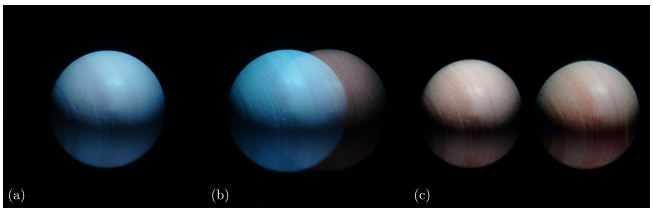


Fig. 13. (a) Photograph of a virtual object displayed on top of the real object. We changed the hue of the texture to show a mixed-reality effect. (b) The real object can be seen more clearly with the displayed object slightly shifted away. (c) Photograph of the displayed object (right) next to the real object (left). The small white strip visible on the bottom right corner of the right object is not a display artifact but reflection from the background.

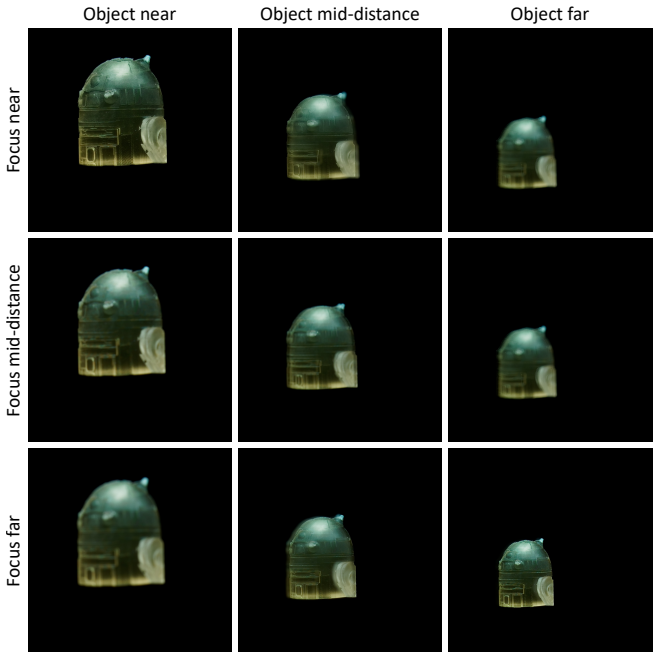


Fig. 14. Photographs of an object rendered on our display at different depths (columns) while the camera focus was set to one of the three fixed focal distances (rows). The photographs demonstrate the performance of defocus blur due to the multi-focal plane rendering. Note that the subpixel structure, seen in magnification, is not noticeable when the object is seen by the eye. The position of the object changes in the field of view since the camera optical axis was not aligned with the object depth axis.

Figure 14 shows photographs of a rendered 3D-printed robot figure, displayed at three distances while the camera was set to one of those three focal distances. As expected, the display shows a desired defocus blur when the object is shown at a different focal depth from that of the camera lens. However, since there is no display focal plane in the mid-distance, the image shown at the center is a superimposition of the two defocused images from both focal planes, which results in a visually incorrect blur. The amount of such blur can be reduced by bringing both focal planes closer.

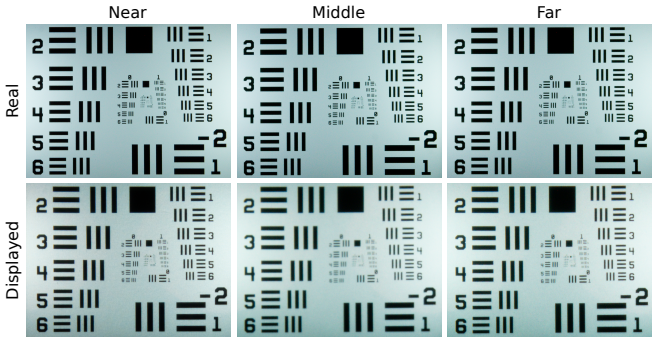


Fig. 15. Photographs of a physical 2D resolution chart (top) in comparison with its displayed counterpart (bottom) placed at different depths (columns).

To evaluate the resolution limit and the aforementioned incorrect defocus blur (when the virtual object is placed between the two focal planes) of our display, we reproduced a 1951 USAF resolution test chart (ThorLabs, R3L3S1P, positive, 3"×3") and photographed it in comparison with the physical chart (Figure 15). We built a custom light box to illuminate the chart from the back, producing a high-contrast resolution pattern. We displayed either the real or rendered virtual chart at one of three distances[1]: 500 mm (near), 577 mm (middle), and 654 mm (far). The camera focus was also set to one of these distances. To reduce the Moiré pattern resulting from the interference of the LCD and camera sensor pixel grids, we reduced the aperture to F16 and processed the images using *DxO PhotoLab 4.3.0* with only Moiré filtering enabled. Note that the Moiré pattern was not visible to naked eyes. Assuming that the resolution limit is the point at which the lines blend together and cannot be regarded as separate, our display can reproduce up to 4.0 lp/mm at 500 mm (0.58 lp/arcmin), 2.83 lp/mm at 577 mm (0.48 lp/arcmin) and 4.0 lp/mm at 654 mm (0.76 lp/arcmin). This shows a dip for the middle distance, at which the displayed image is a superimposition of two defocused focal planes (Figure 15, 2nd row, 2nd column).

6 EXPERIMENT: VISUAL TURING TEST

We designed an experiment to test whether participants can distinguish between real and virtual objects shown by our system. The experiment is inspired by the early work of Meyer et al. [1986] and many follow-up studies, which attempted to create a system that passes a *computer graphics Turing test* or *virtual reality visual Turing test*. In contrast to these studies, which have reproduced only 2D images of limited dynamic range, we have created a capture-and-display system that can deliver all necessary visual cues. The secondary objective of our experiment is to test the sensitivity of the visual system to the degradation of different cues (contrast, in this experiment) when all other cues are present. We hope that such data will facilitate understanding of what trade-offs are acceptable in the fidelity of individual display properties, while still delivering highly realistic content — valuable information for building practical display systems.

[1]For this evaluation, we moved the near focal plane close to the near distance and far focal plane close to the far distance.

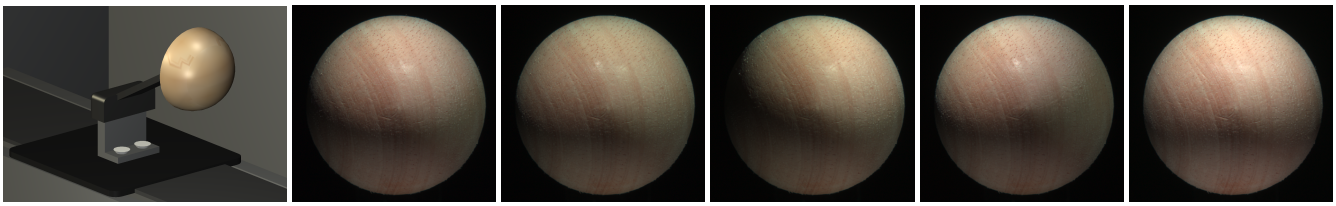


Fig. 16. The 3D CAD model of the object (left) and its photographs under the five illumination patterns used in the experiment. The base of the wooden hemisphere had a diameter of 47 mm. The photographs have been tone-mapped with $\gamma = 2.2$ to preserve the original colors.

Stimuli. Our test object was a wooden hemisphere (a prop used to teach geometry) that was lightly sanded and stained, but retained the texture of wood and produced an imperfect specular reflection of moderate intensity (refer to Figure 16). As shown on the left of Figure 16, the hemisphere was attached to a 3D printed holder (504 mm from the viewer) on the flat side and had its spherical side directed toward the viewer so that it appeared as a sphere to a participant. We selected this object for its simple geometry and complex material and texture properties.

The sphere was illuminated by one of five different light patterns, produced by the RGB LED array on the ceiling of the real-scene box. The patterns were created by switching on a set of 2 LEDs at different positions in the LED array so that the object was illuminated from a slightly different angle each time (while keeping overall brightness approximately the same). To indirectly illuminate the object from the bottom, a piece of white cardboard acting as a diffuse reflector was placed under the object. Different illumination patterns are an important part of our experiment design as they let vary the stimulus between the trials so that the participants cannot not memorize small differences in appearance across the trials.

A rectangular aperture, made of black cardboard, was placed on the front side of the real-scene box so that only the illuminated hemisphere can be seen. The illumination was reduced to the point at which only the hemisphere can be seen but not any part of the real-scene box (the peak luminance of the object was 2 cd/m^2).

In addition to the *standard* condition, which was our best reproduction of the real object, we created a distorted condition, in which we artificially reduced contrast. The contrast was reduced by modifying pixel values:

$$I_{\text{mod}}(x, y, c) = \left(\frac{I_{\text{org}}(x, y, c)}{I_{\text{med}}} \right)^{\gamma} I_{\text{med}}\,, \tag{4}$$

where I_{med} is the median luminance of the image, I_{org} and I_{mod} are the original and modified images (in linear RGB color space), and (x, y, c) are pixel and color channel indices. We determined in a pilot experiment that $\gamma = 0.8$ produced results that were detectable but sufficiently challenging. Not only does this condition let us evaluate the effects of reducing contrast per se, but it also plays important role in our experiment design that it allows us to exclude the possibility that the task given to the participants was too difficult to be feasible (or that they are not paying adequate attention). Consider the case where we reduce presentation time, or luminance, such that none of the participants can detect the real stimulus amongst rendered alternatives. This pattern of data would resemble passing the visual Turing test, but for an entirely trivial reason. Showing that people *can* detect small reductions in contrast with our chosen experiment parameters, however, would demonstrate that they did perform the discrimination task satisfactorily, and so a failure to discriminate in the standard condition can be interpreted at face value.

The object was rendered either on the near focal plane of our display or on both focal planes and using linear blending in diopter space, as explained in Section 4.4. We tested both conditions to understand the importance and also challenges of delivering correct focal depth.

Procedure. We used a three-interval-forced-choice (3IFC), or odd-one-out, procedure. In each trial, the participant was shown three intervals, for 2 seconds each, from which either two were real and one virtual, or two were virtual and one real. The participant was given the instruction: *You will see three objects, one after another. Select the object that appears different from the two others.* We intentionally avoided asking a question about realism as such a question would be open to subjective interpretations of what "real" looks like, and may lead observers to attend to some aspects of the stimulus while ignoring others. With an oddity task, the observer was instead free to use any aspect of the stimulus to make their judgement, making it a true test of the ability to discriminate real from rendered images. Indeed, the 3IFC task can be considered a very strict test of our display, given that in practical use observers will often evaluate the realism of a rendered scene without the presence of an equivalent real comparison. To avoid after-images causing identical stimuli to appear different between intervals, we showed a plane with a noise texture of the same average luminance as the object and at the same distance.Our procedure aims to objectively measure whether observers are able to distinguish a real object from a virtual one without being provided any training, prior knowledge or experience for the given task.

The experimental session consisted of 120 trials, which took on average 40 minutes to complete, split into two sessions with a short break. Each participant completed 30 repetitions of each condition. In each trial, we randomly selected either a standard stimulus or one with reduced contrast condition, and presented it using either 2-focal plane rendering, or only on the front focal plane (4 conditions in total). One of five illumination patterns was randomly selected for each trial (the same pattern was used in all three intervals). As the alignment of two focal planes is crucial for the reproduction of focal distance, we displayed an alignment grid (similar to Figure 11) before each trial. The participants pressed a key to continue only when good alignment was achieved. They also had an option to repeat the trial if they were distracted or accidentally moved their

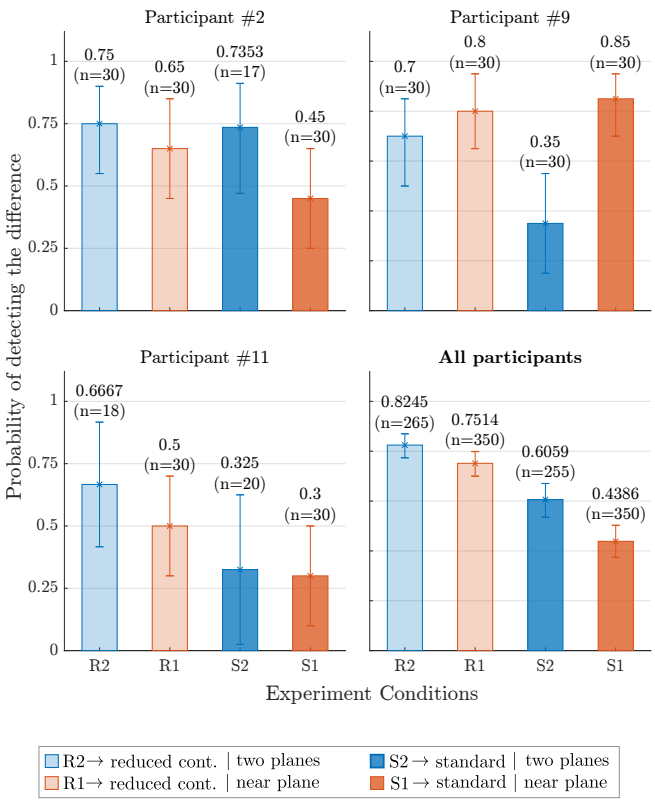


Fig. 17. The probability of detecting for each condition (compensated for the guess rate). The results are shown for three selected participants and averaged across all participants. The results of each individual participant can be found in the supplementary materials.

heads. Finally, we asked the participants to wear glasses frames with an IR-LED (Figure 6), which was used to track and record their head position before and after each interval. We removed the measurements for the trials in which the movement reported by the head tracking was above certain threshold while multi-focal rendering was used (≈ 15% of the measurements).

Participants. 12 participants (three females) completed the experiment. The age of participants ranged from 23 to 34 with the mean of 27.8. Each participant was screened for normal stereo acuity with the *Titmus fly test* and for normal color vision with the *Ishihara test*. The participants were instructed to wear their corrective optics. They were compensated for their participation.

Results. The participants' answers give us a measure of probability of selecting the correct interval, $P(\text{correct})$. Since the participants can select the correct answer by chance, we need to correct for that by modeling:

$$\begin{aligned} P(\text{correct}) &= P(\text{chance} \cup \text{detected}) \\ &= P(\text{chance}) + P(\text{detected}) - P(\text{chance})P(\text{detected}), \end{aligned} \tag{5}$$

where $P(\text{chance}) = 1/3$ in a 3IFC experiment. $P(\text{detected})$ does not depend on the protocol (2IFC or 3IFC) and a zero $P(\text{detected})$ indicates a complete perceptual equivalence between the real and virtual objects. The resulting probability of detecting the interval that appears different, $P(\text{detected})$, is plotted in Figure 17 for three selected participants and also averaged across all 12 participants. As expected, the results show that the reduced contrast increases the probability of detecting the different object, proving that the participants can perform the task. However, multi-focal rendering on both planes made it easier to perform the task compared to rendering only to the near plane (with the exception of participant #9, see Figure 17). We discuss potential factors that contribute to this outcome in Section 7.

Table 1. The results of the post-experiment questionnaire in which the participants were asked to tick one or more differences they could see between the real and virtual objects.

Options	Votes
different color	*2*
different sharpness	*6*
different brightness	*4*
different shape or size	*0*
different position or orientation	*0*
different illumination	*1*
one object or other objects appeared flatter	*0*
one object or other objects appeared less shiny	*6*
material appeared different	*1*

The results also show large individual differences in detection probabilities across participants (see the supplementary materials for all individual results). This is most likely because different participants tend to pay attention to different aspects of the stimuli. We collected a post-experiment questionnaire to better understand how the participants attempted to identify the different object. In the questionnaire, we asked: *What made the selected object stand out from the other objects?* and gave a set of possible answers listed in Table 1. Table 1 shows that among the 12 participants, six participants ticked *sharpness*, which could be a result of the incorrect defocus blur discussed in Section 5 or the insufficient resolution (compared to human sensitivity) of our display. The option *one object or other objects appeared less shiny* was also ticked by six participants. This is potentially due to an inaccuracy of our lumigraph synthesis approach, since the shininess of an object is attributed to specular reflections. Four participants selected *brightness* while two selected *color*, indicating room for improvements in our photometric calibration and color reproduction. We elaborate on the aforementioned issues in Section 7. All participants reported that none of the virtual stimuli appeared unnatural when viewed in isolation and if they had not been asked to look for differences from a physical stimulus, they would have deemed the virtual stimuli to be real.

We use our measurements across 4 conditions to further isolate the factors that contributed to the detection. Assuming that all factors are independent but multiple factors can trigger the detection, we can model the probability of detection as the probability summation:

$$P(\text{detected}) = 1 - (1 - P(\text{f1}))\,(1 - P(\text{f2}))(1 - P(\text{contrast})), \tag{6}$$

where $P(\text{f1})$ is the probability of detecting the difference due to single focal plane rendering, $P(\text{f2})$ is the probability of detecting the artifacts to the limitations of two-focal plane rendering (excluding all factors contributing to $P(\text{f1})$) and $P(\text{contrast})$ is the probability of detecting reduced contrast. We use maximum likelihood estimation to compute those probabilities across all participants and get:

$$P(\text{f1}) = 0.44 \quad P(\text{f2}) = 0.3 \quad P(\text{contrast}) = 0.56\,. \tag{7}$$

This shows the observers have 44% chance of detecting the difference between real and virtual objects shown by our display and that two-focal plane rendering increases that chance by 30%[2]. The isolated probability of detecting the contrast reduction by 20% ($\gamma = 0.8$) is 56%, which corresponds to about 1 JND unit (78% for a 2IFC protocol). The reduced contrast conditions serve as an example of a procedure that can be used to scale other relevant "distortions", such as the change of luminance, disparity or black level.

7 DISCUSSION

3IFC task. The outcome of our experiment, showing that observers can detect the virtual object in 44% of the cases may appear worse than the results reported in other works [Borg et al. 2012; Meyer 1986]. However, we need to consider that this is the first time a direct comparison was made between a display and a 3D object seen from a short distance. We also used a much more challenging 3IFC procedure, which removed the subjective assessment of "realism" from our task, and made our test sensitive to very small differences between displayed and real objects. Such differences in certain insignificant aspects (such as viewing angle, object size, position, etc.) do not necessarily degrade the quality of realism for images viewed in isolation.

Distorted conditions. Most visual experiments in graphics either test preference (does A look better than B) or measure similarity to a "reference", which is often obtained from costly rendering, such as path-tracing. Both approaches can only be used to determine relative improvements with regard to another rendering method, which may or may not capture the desired visual qualities. Our reduced contrast condition demonstrated how a (simulated) rendering method (or a display limitation) can be directly compared against the ultimate reference of a real-world object. Such absolute measures can tell us that a certain percentage of observers across a population will not notice any observable difference to the real-world object ($P(\text{contrast})$), while discounting the existing imperfections of the display ($P(\text{f1})$ and $P(\text{f2})$). We plan to use such a methodology to quantify the importance of various display capabilities, such as the dynamic range, absolute luminance, disparity, focal distance, accommodation, and others.

Eye tracking. Multi-focal rendering requires very precise alignment across the focal planes. Effective alignment without uncomfortable restraining of head position requires active tracking and compensation for the head position. Our IR LED tracker was a first step toward this goal. Latency of the tracking, and limited refresh-rate of the display, did not let us implement active compensation for head movement yet. These are not fundamental limitations of the approach, however.

Multi-focal rendering. Our experiment showed a result that rendering on two planes with linear depth filtering made it easier for most observers to detect discrepancy. One explanation could be that while linear depth filtering with the current two-plane separation distance can drive accommodation to the correct depth, it causes an increased defocus blur compared to real scenes. Any multi-focal-plane display with a practical number of focal planes necessarily samples focal depth coarsely, and so most scene points will not coincide precisely with a focal plane. Accommodation can be driven to the appropriate inter-plane distances by linear depth filtering (with plane separations up to and even exceeding that used here, [MacKenzie et al. 2010]). Yet, at least one image plane must be defocused (because two cannot be focused on simultaneously), resulting in potentially detectable blur compared to a real scene. The results suggest defocus blur plays a more important role in perceptual realism than the accommodation response. As we are relatively insensitive to accommodation state, and it is a weak depth cue, incorrect accommodation is likely to provide weaker cues to realism than blur. Several steps can be taken, however, to reduce this defocus blur compared to the present study. Due to light scattering inside the real-scene box, we used dim illumination, which increased the pupil size, thereby increasing defocus blur. In rendered scenes this problem can be reduced by using higher luminance (including HDR) scenes. Also, in this study the far focal plane was quite far from the stimulus. Either adding an additional intermediate focal plane, or moving the planes to optimal positions with respect to the scene content, would reduce the focal depth inaccuracies that lead to increased defocus blur. Finally, more advanced multi-focal decomposition algorithms may be able to compensate for the loss of high spatial frequencies that characterizes defocus blur [Mercier et al. 2017; Narain et al. 2015]. Since a correct stimulus to accommodation is necessary to avoid problems caused by vergence-accommodation conflicts [Hoffman et al. 2008], there is great value in attempting to optimize multi-focal displays for reproducing realism. We hope that our display can be used to explore the trade-offs involved in doing this. For example, does tolerance to incorrect focal depth increase if other aspects of the scene are delivered with very high fidelity?

Reproducible stimuli. Our system has several limitations in terms of the stimuli it can reproduce. While our system can synthesize non-Lambertian materials with specular reflections, the quality may not reach the level of perceptual equivalence, as indicated by our post-experiment questionnaire. Specular highlights are sensitively dependent on viewing positions, making them difficult to be reproduced as it is unlikely that our data camera perfectly overlaps with the observer's eye position. We anticipate that such inaccuracies can be reduced by capturing more light field views or incorporating advanced neural scene representations such as neural radiance fields [Mildenhall et al. 2020] or view-dependent multi-plane images [Wizadwongsa et al. 2021]. We did not explore this direction as training and convergence of scene-representation networks with large-size data (8k images in our case) remains an actively studied problem. In the future, we plan to evaluate various view synthesis approaches with our apparatus in terms of perceptual realism,

[2]Note that the probability of detecting limitations of single focal plane or two focal plane rendering (or both) is: $P(\text{f1} \cup \text{f2}) = P(\text{f1}) + P(\text{f2}) - P(\text{f1})P(\text{f2}) = 0.61$.

whereas existing works only focus on photorealism. Our system is also currently limited to simple or known geometry. Nonetheless, this is not a fundamental limitation of our approach — we can modify the loss function in Eq. (1) such that it not only optimizes for a spatial transform but also for a per-vertex deformation to fit an unknown geometry. However, this approach also requires capturing many more views around the object. Since 3D reconstruction is not the main focus of this work, we chose to work with known geometry and a horizontal light field. Our rendering method is currently unable to reproduce edge occlusions of objects at different depths without introducing visible artifacts. Our intention is to test more advanced multi-plane rendering methods [Mercier et al. 2017; Narain et al. 2015; Yu et al. 2019] in the future. Our display has an advantage over the previously built multi-focal plane displays in that it can reproduce a much higher dynamic range, which gives more flexibility in optimizing for multi-plane decomposition (for example, greater headroom for compensating for loss of high spatial frequencies). In addition, although the resolution of our display is much higher than that found in the previous work [Vangorp et al. 2014], it is still lower than the levels required for a perceptual match, as reported by Masaoka et al. [2013]. Achieving the highest resolution reported in their paper (120 cpd) would require tripling the resolution of our LCD panels. This is currently impossible when using off-the-shelf components. As above, it will be interesting to explore whether tolerance to lower-than-optimal resolution is increased when other aspects of the scene are delivered with high fidelity. Finally, our color calibration currently relies on CIE XYZ 1931 color matching functions, which are known to be inaccurate for short wavelengths [CIE170-1:2006 2016]. It also did not account for the contribution of rods to color perception or individual differences. Better color matching may require capturing multispectral images and individual corrections to compensate for the differences in cone sensitivities.

8 CONCLUSIONS

The main objective of our work is to build an end-to-end system that can acquire a small 3-dimensional object and reproduce it faithfully with all the necessary visual cues on a display. Being able to do so is an important step for perceptually realistic graphics, in which the depicted imagery is indistinguishable from the real world. A direct comparison with real-world objects lets us better understand the limitations of not only the visual system but also those of display technologies, 3D representations, and rendering techniques. For example, we found that defocus blur could play a more important role than accommodation response in perceptual realism, together with the need for accurate view-point tracking, as one of the main limitations of multi-focal plane displays.

We demonstrate that the first iteration of our HDR-MF-S display can deliver virtual imagery that is in only 44% of the cases detected as different from its real-world counterpart. This result was obtained when asking the question "is it different?" rather than "is it real?", making the task more objective but also requiring higher accuracy from a display system. We believe this is also the first attempt to reproduce a 3D object at a short distance, with the essential set of depth cues for the given scene. Finally, our experiment design with a "control" distorted condition (reduced contrast) ensured that the participants were correctly completing the task.

The display is a platform for a wide range of experimental studies, in which both faithful reproductions of all visual cues and comparison to reality are paramount. For example, it can be used for studies on gloss and material perception, physics-based rendering, global illumination, tone mapping, view synthesis, augmented & mixed reality, and many more. All these studies can take advantage of full control over each display capability dimension, such as dynamic range or luminance. The displays can also simulate a wide range of see-through AR displays, by using a real-scene box as a real environment and offering a much higher dynamic range and peak luminance than that of most head-mounted displays.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the European Research Council (ERC) Consolidator Grant agreement N° 725253 (EyeCode) and under the Marie Skłodowska-Curie grant agreement N° 765911 (RealVision).

REFERENCES

Kurt Akeley, Simon J. Watt, Ahna Reza Girshick, and Martin S. Banks. 2004. A Stereo Display Prototype with Multiple Focal Distances. ACM Trans. Graph. 23, 3 (Aug. 2004), 804–813. https://doi.org/10.1145/1015706.1015804

AliceVision. 2018. Meshroom: A 3D reconstruction software. https://github.com/alicevision/meshroom

Martin S. Banks, David M. Hoffman, Joohwan Kim, and Gordon Wetzstein. 2016. 3D Displays. Annual Review of Vision Science 2, 1 (2016), 397–435. https://doi.org/10.1146/annurev-vision-082114-035800 arXiv:https://doi.org/10.1146/annurev-vision-082114-035800 PMID: 28532351.

M. Borg, S. S. Johansen, D. L. Thomsen, and M. Kraus. 2012. Practical Implementation of a Graphics Turing Test. In Advances in Visual Computing, George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Charless Fowlkes, Sen Wang, Min-Hyung Choi, Stephan Mantler, Jürgen Schulze, Daniel Acevedo, Klaus Mueller, and Michael Papka (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 305–313. https://doi.org/10.1007/978-3-642-33191-6_30

Jen-Hao Rick Chang, B. V. K. Vijaya Kumar, and Aswin C. Sankaranarayanan. 2018. Towards Multifocal Displays with Dense Focal Stacks. ACM Trans. Graph. 37, 6, Article 198 (Dec. 2018), 13 pages. https://doi.org/10.1145/3272127.3275015

CIE170-1:2006. 2016. Fundamental chromacity diagram with psychological axes - part 1. Technical Report. Central Bureau of the Commission Internationale de l' Éclairage. http://www.cie.co.at/publications/fundamental-chromaticity-diagram-physiological-axes-part-1

David Dunn, Cary Tippets, Kent Torell, Petr Kellnhofer, Kaan Akşit, Piotr Didyk, Karol Myszkowski, David Luebke, and Henry Fuchs. 2017. Wide Field Of View Varifocal Near-Eye Display Using See-Through Deformable Membrane Mirrors. IEEE Transactions on Visualization and Computer Graphics 23, 4 (2017), 1322–1331. https://doi.org/10.1109/TVCG.2017.2657058

G. Eilertsen, R. K. Mantiuk, and J. Unger. 2017. A comparative review of tone-mapping algorithms for high dynamic range video. Computer Graphics Forum 36, 2 (may 2017), 565–592. https://doi.org/10.1111/cgf.13148

Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. 1984. Modeling the interaction of light between diffuse surfaces. ACM SIGGRAPH Computer Graphics 18, 3 (jul 1984), 213–222. https://doi.org/10.1145/964965.808601

Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. 1996. The lumigraph. In Proc. of SIGGRAPH '96. ACM Press, 43–54. https://doi.org/10.1145/237170.237200 arXiv:0070242542

Param Hanji, Fangcheng Zhong, and Rafał K. Mantiuk. 2020. Noise-Aware Merging of High Dynamic Range Image Stacks without Camera Calibration. In Advances in Image Manipulation (ECCV workshop). Springer, 376–391. http://www.cl.cam.ac.uk/research/rainbow/projects/noise-aware-merging/

David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks. 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. Journal of Vision 8, 3 (mar 2008), 33. https://doi.org/10.1167/8.3.33

Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. 2015. The Light Field Stereoscope: Immersive Computer Graphics via Factored near-Eye Light Field Displays with

Focus Cues. ACM Trans. Graph. 34, 4, Article 60 (July 2015), 12 pages. https://doi.org/10.1145/2766922

Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. 2019. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. The IEEE International Conference on Computer Vision (ICCV) (Oct 2019). https://arxiv.org/abs/1904.01786

Mark E. Lucente. 2012. Electronic Holographic Displays: 20 Years of Interactive Spatial Imaging. In Handbook of Visual Display Technology, Janglin Chen, Wayne Cranton, and Mark Fihn (Eds.). Springer-Verlag Berlin Heidelberg, Bristol, 2721–2740.

Kevin J. MacKenzie, Ruth A. Dickson, and Simon J. Watt. 2012. Vergence and accommodation to multiple-image-plane stereoscopic displays: "real world" responses with practical image-plane separations? Journal of Electronic Imaging 21, 1 (feb 2012), 011002. https://doi.org/10.1117/1.JEI.21.1.011002

Kevin J MacKenzie, David M Hoffman, and Simon J Watt. 2010. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. Journal of vision 10, 8 (jan 2010), 22. https://doi.org/10.1167/10.8.22

K. Masaoka, Y. Nishida, M. Sugawara, E. Nakasu, and Y. Nojiri. 2013. Sensation of Realness From High-Resolution Images of Real Objects. IEEE Transactions on Broadcasting 59, 1 (mar 2013), 72–83. https://doi.org/10.1109/TBC.2012.2232491

Ann. M. McNamara. 2005. Exploring perceptual equivalence between real and simulated imagery. In Proceedings of the 2nd symposium on Appied perception in graphics and visualization - APGV '05. ACM Press, New York, New York, USA, 123–128. https://doi.org/10.1145/1080402.1080425

Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. 2006. Demosaicing with directional filtering and a posteriori decision. IEEE Transactions on Image Processing 16, 1 (2006), 132–141. https://doi.org/10.1109/TIP.2006.884928

Olivier Mercier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. 2017. Fast Gaze-Contingent Optimal Decompositions for Multifocal Displays. ACM Trans. Graph. 36, 6, Article 237 (Nov. 2017), 15 pages. https://doi.org/10.1145/3130800.3130846

G. W. Meyer. 1986. An experimental evaluation of computer graphics imagery. ACM Transactions on Graphics 5, 1 (jan 1986), 30–50. https://doi.org/10.1145/7529.7920

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In ECCV.

Rahul Narain, Rachel A. Albert, Abdullah Bulbul, Gregory J. Ward, Martin S. Banks, and James F. O'Brien. 2015. Optimal Presentation of Imagery with Focus Cues on Multi-Plane Displays. 34, 4, Article 59 (July 2015), 12 pages. https://doi.org/10.1145/2766909

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3D Deep Learning with PyTorch3D. arXiv:2007.08501 (2020).

Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. 2002. Photographic tone reproduction for digital images. ACM Transactions on Graphics 21, 3 (jul 2002), 267. https://doi.org/10.1145/566654.566575

T Ritschel, M Ihrke, J. R. Frisvad, J. Coppens, K. Myszkowski, and H.-P. Seidel. 2009. Temporal Glare: Real-Time Dynamic Simulation of the Scattering in the Human Eye. Computer Graphics Forum 28, 2 (apr 2009), 183–192. https://doi.org/10.1111/j.1467-8659.2009.01357.x

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": Interactive Foreground Extraction Using Iterated Graph Cuts. ACM Trans. Graph. 23, 3 (Aug. 2004), 309–314. https://doi.org/10.1145/1015706.1015720

Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. 2004. High Dynamic Range Display Systems. ACM Trans. Graph. 23, 3 (Aug. 2004), 760–768. https://doi.org/10.1145/1015706.1015797

Phil Surman and Ian Sexton. 2012. Emerging Autostereoscopic Displays. In Handbook of Visual Display Technology, Janglin Chen, Wayne Cranton, and Mark Fihn (Eds.). Springer-Verlag Berlin Heidelberg, Bristol, 2652–2667.

I.E. Sutherland. 1974. Three-dimensional data input by tablet. Proc. IEEE 62, 4 (1974), 453–461. https://doi.org/10.1109/PROC.1974.9449

Peter Vangorp, Rafat K Mantiuk, Bartosz Bazyluk, Karol Myszkowski, Radosław Mantiuk, Simon J Watt, and Hans-Peter Seidel. 2014. Depth from HDR: depth induction or increased realism?. In ACM Symposium on Applied Perception - SAP '14. ACM Press, New York, New York, USA, 71–78. https://doi.org/10.1145/2628257.2628258

Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. 2021. NeX: Real-time View Synthesis with Neural Basis Expansion. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Akiko Yoshida, Rafał Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. 2006. Analysis of Reproducing Real-World Appearance on Displays of Varying Dynamic Range. Computer Graphics Forum (Proc. of Eurographics) 25, 3 (sep 2006), 415–426. https://doi.org/10.1111/j.1467-8659.2006.00961.x

Hyeonseung Yu, Mojtaba Bemana, Marek Wernikowski, Michał Chwesiuk, Okan Tursun, Gurprit Singh, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, and Piotr Didyk. 2019. A Perception-driven Hybrid Decomposition for Multi-layer Accommodative Displays. IEEE transactions on visualization and computer graphics PP (02 2019). https://doi.org/10.1109/TVCG.2019.2898821