Off-Axis Layered Displays: Hybrid Direct-View/Near-Eye Mixed Reality with Focus Cues

Christoph Ebner, Peter Mohr, Tobias Langlotz, Yifan Peng, Dieter Schmalstieg, Gordon Wetzstein, Denis Kalkofen

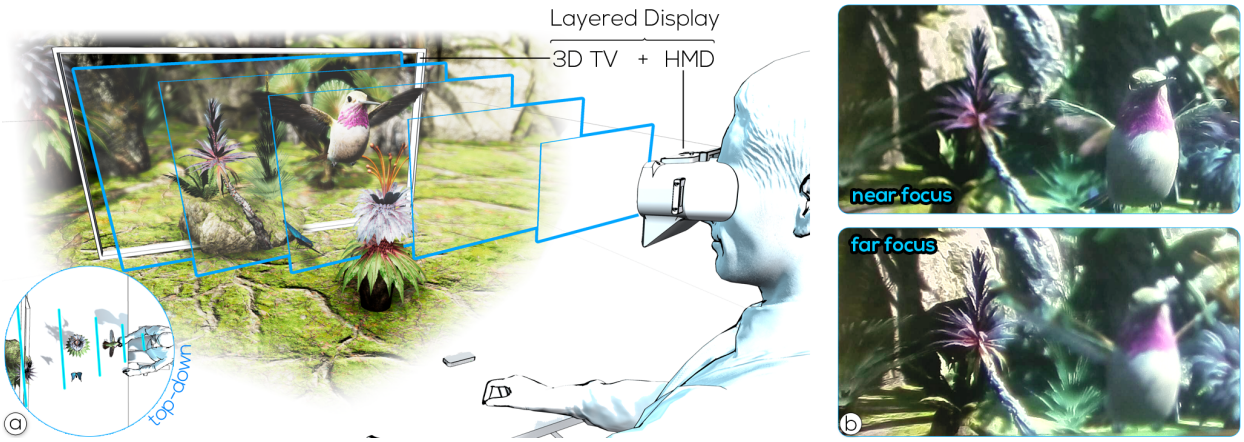

Fig. 1. Overview of the concept of off-axis layered displays. (a) Off-axis layered displays combine a direct-view display, such as a desktop monitor, projector, or handheld display, with a near-eye display, to provide focus cues. Input to off-axis layered displays is a focal stack of the 3D scene. The focal stack is illustrated between the user and the direct-view display. (b) Photographs captured through the HMD for demonstrating the capability of the display to provide in-focus contrast at different focus distances.

Abstract— This work introduces off-axis layered displays, the first approach to stereoscopic direct-view displays with support for focus cues. Off-axis layered displays combine a head-mounted display with a traditional direct-view display for encoding a focal stack and thus, for providing focus cues. To explore the novel display architecture, we present a complete processing pipeline for the real-time computation and post-render warping of off-axis display patterns. In addition, we build two prototypes using a head-mounted display in combination with a stereoscopic direct-view display, and a more widely available monoscopic direct-view display. In addition we show how extending off-axis layered displays with an attenuation layer and with eye-tracking can improve image quality. We thoroughly analyze each component in a technical evaluation and present examples captured through our prototypes.

Index Terms—Layered display, Direct-view, Near-eye, Mixed Reality, Focus cues

1 Introduction

Information displays, particularly two-dimensional (2D) direct-view displays, have become ubiquitous in everyday life. Examples include PC monitors at work, TVs or large video wall panels for entertainment, and handheld displays for mobile visualization applications. Three-dimensional (3D) displays extend on traditional 2D screens by leveraging additional depth cues [18]. Most commonly, 3D displays support stereoscopic viewing by showing two slightly different images to the user's left and right eye for delivering binocular disparity.

Current-generation direct-view 3D displays, however, offer only a subset of the perceptually important depth cues that humans encounter in real-world environments. For instance, common 3D displays suffer from the vergence–accommodation conflict (VAC) as they provide binocular disparity on a single, fixed focal plane only. Recent advances in near-eye displays in virtual reality (VR) and augmented reality (AR) systems mitigate the VAC by shifting the focal plane based on eye-tracking data [26], by using holographic display solutions [7, 46], or using multiple display layers [14]. Built upon these approaches, near-eye displays can provide near-correct focus cues, and thus provide perceptually realistic and visually comfortable user experiences.

Unfortunately, the image quality of AR and VR displays cannot yet compete with that of traditional screens. For example, modern direct-view displays provide much higher resolutions, measured in pixels per visual degree, than any commercially available head-mounted display (HMD). Moreover, sharing a VR experience with other users in the same physical space is challenging, potentially resulting in users feeling more isolated and increasing the difficulty of mundane tasks, such as typing or drinking [5, 37].

To address these challenges, we choose to extend the capabilities of traditional direct-view displays. Specifically, we introduce the concept and prototype for *off-axis layered displays*, which combine direct-view and near-eye displays to form a viewing volume with support for focus cues to improve perceptual realism and visual comfort. Our approach is built on the insight that a conventional display can be extended to a layered display by placing one or more optical see-through head-worn display layers in front of it (see Figure 1). As such, users of our system are able to perceive their surroundings unobstructed but they see also

- *Christoph Ebner, Peter Mohr, and Dieter Schmalstieg are with Graz University of Technology. E-mail: christoph.ebner@icg.tugraz.at*
- *Tobias Langlotz is with the University of Otago, E-mail: tobias.langlotz@otago.ac.nz*
- *Yifan Peng is with the University of Hong Kong, E-mail: evanpeng@hku.hk*
- *Gordon Wetzstein is with Stanford University, E-mail: gordonwz@stanford.edu.*
- *Denis Kalkofen is with Flinders University and Graz University of Technology. E-mail: kalkofen@icg.tugraz.at*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

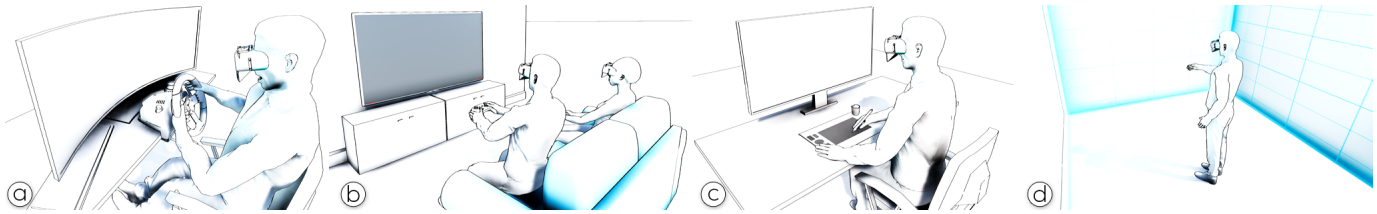

Fig. 2. Applications. Various professional and casual applications can benefit from enhanced visual perception by extending existing direct-view displays into off-axis layered displays. Example use-cases include, (a) single- and (b) multi-user gaming at home, (c) stereoscopic viewing during CAD modelling on a desktop, and (d) exploration and interaction in immersive projector environments, such as a CAVE.

the 3D content when directly viewing the direct-view display allowing for more natural collaboration and interaction with the environment. Although extending 2D screens into a layered 3D display has been reported for auto-stereoscopic viewing before [55], we present the first approach that considers a hybrid direct-view/near-eye display setup that continuously updates the spatial arrangement between the stationary and wearable component.

Focus cues must be provided in a large enough depth interval, but – most critically – within the user's direct interaction space [10], i.e., up to approximately 1.5-2 m. Thus, we design off-axis layered displays to provide focus cues in the user's close proximity. Our approach combines an optical see-through HMD with a direct-view display to set up a layered display within the user's direct interactive space (Figure 1). The wearable display plane is placed at 3 diopter (dpt) in front of the user, which enables focus cues from approximately 30 cm to where the direct-view display is placed, i.e., approximately at a distance of 1.5-2 m in a common setting. Note that the size of the view volume interactively adjusts based on the distance between the user and the direct-view displays. The combination of one stationary and one head-worn display layer lends itself to scenarios where existing display technology can be re-purposed. Users purchasing a see-through HMD may have access to large format direct-view displays in typical office environments or homes. If these two independent displays are synchronized, they can be used as an off-axis layered display for 3D viewing with focus cues, which the HMD or the stationary display cannot provide in standalone mode. Hence, opportunities arise for various professional and casual applications benefiting from enhanced visual perception. Examples include, authoring and viewing of geometric and CAD data, spatial telepresence, immersive movies, and games and simulations in desktop, home and CAVE like environments. The illustrations in Figure 2 depict several configurations and applications.

Off-axis layered displays have been designed to mitigate the VAC that is common in traditional stereoscopic displays. However, a layered display that consists of wearable and direct-view displays offers several other unique characteristics. For example, off-axis layered displays allow multiple users to interact with a shared direct-view display while simultaneously being able to view individual perspectively correct per-user content (Figure 2b). High-resolution light fields can be displayed with a color fidelity matching those of 2D displays. Moreover, our display can switch between 2D and 3D for applications that require working on 2D documents and 3D objects simultaneously.

To explore off-axis layered displays, we discuss key challenges and show how they can be addressed in a working prototype. In particular, we introduce a novel decomposition method, which we use to generate display patterns for a dynamically skewed viewing volume between a stationary and a wearable display layer, by decomposing a focal stack of input images. Additionally, we introduce approaches for post-render warping of the patterns, and achieve high in-focus contrast using eye-tracking data. Hence, our work makes the following contributions:

- We introduce off-axis layered displays, a novel class of interactive 3D displays that can deliver focus cues within a volume between a user who is wearing an HMD and a 2D or 3D direct-view display.
- We build two prototypes of off-axis layered displays using an HMD in combination with (I) a stereoscopic direct-view display, and (II) a more widely available monoscopic direct-view display. Therefore, we develop novel approaches for the real-time computation and post-render warping of off-axis display patterns.
- We present technical extensions to off-axis layered displays. Specifically, we introduce an attenuation layer, we extend the layer decomposition to multiple users, and we demonstrate how eye-tracking can improve the performance of layered displays.
- We analyze off-axis layered displays in a technical evaluation.

2 Related Work

We combine a stationary and a wearable display layer to enable stereoscopic viewing with focus cues. To put this novel display scheme in context, we review previous work on stereoscopic displays, accommodation-supporting displays, and those combining stationary with wearable displays.

2.1 Stereoscopic displays

3D displays enable stereoscopic viewing by delivering a pair of images to its users' left and right eyes [28]. These two images encode binocular depth cues as they are commonly acquired from two horizontally separated perspectives, which mimic the anatomical placement of human eyes. Stereoscopic displays differ in the way how they deliver the image to each eye. Popular approaches rely on shutter glasses for temporal separation [18] and passive glasses with polarization or color filters for spatially separating coded pixels [18]. However, approaches based on color coding suffer from low color fidelity. Temporal approaches reduce the frame rate, and spatial separation reduces the overall number of pixels reaching each eye. Since increasing the frame rate is constrained by the pixel response time, and increasing the resolution is limited by the pixel pitch, both approaches are ultimately bound by physical limitations.

An HMD delivers binocular image pairs also spatially separated. However, by providing a dedicated display panel for each eye, in combination with magnifier lenses used for enlarging the image of a 2D display, an HMD can provide more densely placed pixels to each eye [6]. Still, as an HMD physically places the display panel close to the user's eyes, they commonly provide only a small number of pixels per degree. To overcome the need for wearing glasses, auto-stereoscopic displays directly control the direction of the light that is leaving the display surface [11, 20]. Examples include approaches based on lenticular lenses [18], parallax barriers [18], and liquid crystal layers [27, 54]. While these approaches are able to deliver image pairs without glasses, they do so at the expense of a reduced spatial and angular resolution [36].

In contrast to previous approaches, off-axis layered displays make use of high-resolution screens at arm's length and, therefore, provide more pixels per degree than common HMDs. In addition, off-axis layered displays avoid loss of resolution or frame rate as they are based on unmodified screens and color coding. However, off-axis layered displays support high color fidelity as they optimize pixel colors using tomographic retinal reconstruction.

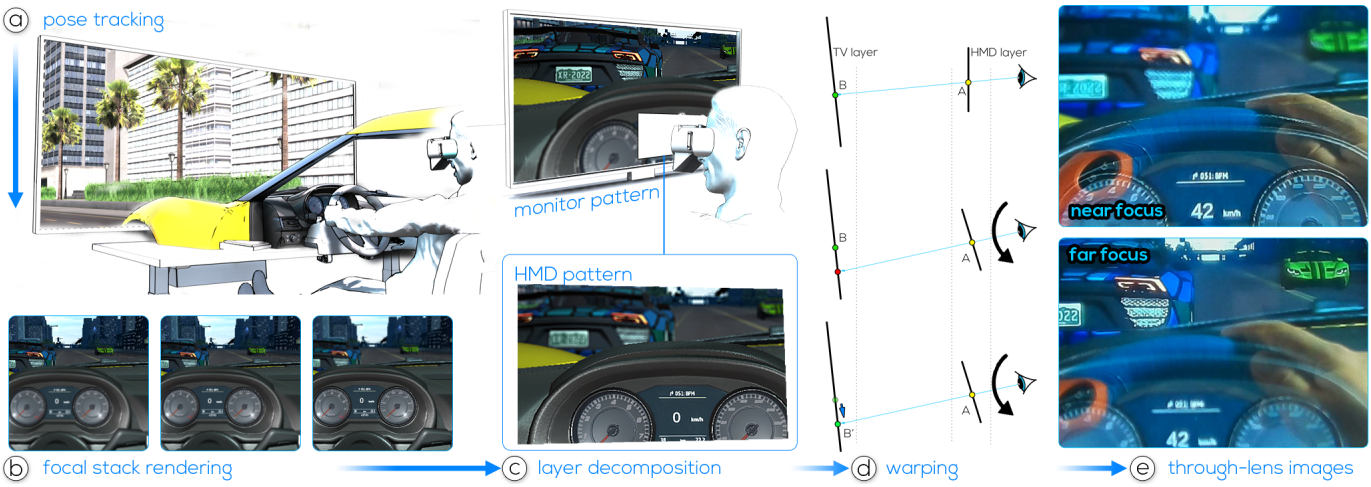

Fig. 3. System overview. (a) We start by estimating the pose of the HMD relative to the direct-view display. (b) For each eye, we render the focal stack, (c) which is subsequently decomposed into image patterns that are displayed on the HMD and the direct-view display. (d) To compensate for rapid user movements, we additionally apply post-decomposition warping before each refresh of the direct-view display. (e) Photographs captured through the HMD demonstrate how the defocus blur adjusts to the focus distance. Notice the change to the contrast of the green car and the speedometer display when they become in focus.

2.2 Accommodation-supporting displays

To mitigate the VAC [19], several approaches have been developed to build displays that provide near-correct focus cues [22, 31]. For example, by modulating the wavefront and reconstructing the wave field of a scene, holographic displays can naturally provide focus cues [23, 44]. With advances in machine learning, the image quality of computer-generated holography has been significantly lifted [9, 45, 49]. However, holographic displays are still in their infancy, which is mostly owed to the limited space-bandwidth product of available spatial light modulators [8, 16]. This limitation along with the costs associated with holographic optical elements still prevents the widespread applicability of interactive holographic displays.

Further designs to mitigate the VAC include varifocal displays, which shift the focus plane to match the users' vergence distance. Designs include mechanical setups [2, 26, 26], deformable mirrors [13], and electrically tunable lenses [35, 48]. These approaches make use of eye tracking to match the distance of the focal plane with the user's vergence distance. However, the accuracy and precision of eye-tracking are often not sufficient to precisely adjust the display system [12]. Therefore, varifocal designs have recently been combined with multifocal approaches [14].

Multifocal displays address several focal planes simultaneously, forming a volume within which near-correct focus cues can be delivered. Common implementations use stacked display layers [34, 55, 57], microlens arrays [32], and high-speed projectors with focus adjusting optics [47, 50]. However, approaches based on high-speed projections with synchronized optics demand complex setups, while approaches based on microlens arrays commonly suffer from the loss of spatial resolution for presenting multi-view images [51].

Therefore, we draw inspiration from layered displays. They require computing a decomposition of the scene into separate image patterns for each layer. Approaches for computing the decomposition use retinal optimization for additive layered displays [38, 42], as well as tomographic reconstruction methods for optimizing combinations of pixel colors of attenuative display layers [51, 55]. Attenuative displays have often been implemented using a stack of transparent liquid crystal displays (LCD)s which commonly suffer from a limited resolution caused by diffraction of light shining through the pixels of an LCD layer [21]. Additive approaches support high-resolution displays but suffer from the lack of blocking light [42]. Hybrid approaches have been presented [25], but only to extend the number of layers. We design the first hybrid layered display, which makes use of high-resolution additive display layers and a low-resolution attenuation layer. This design enables supporting pixel occlusions in high-resolution additive layered displays. Note that a hybrid layered display requires introducing a novel hybrid approach for computing the scene decomposition.

The introduction of display layers arranged in a non-rigid manner leads to off-axis projections which must be resolved at runtime. Thus, we also introduce a new approach for reconstructing the skewed viewing volume between a tracked head-worn display and a direct-view display.

2.3 Hybrid displays

The literature also reports on several examples where different display technologies are combined to create new experiences. For example, FoveatedAR [26] extends the limited field of view of an optical see-through HMD using an external projection into the periphery of the user. This approach is similar to an earlier prototype extending the display area of a direct-view display with projectors [24]. Grubert et al. extended the usable screen space by combining an HMD with a smartwatch or a mobile phone [17]. Similarly, TrackCap [41] combines HMDs capable of displaying 3D information with 2D handheld displays.

More recently, there have been several prototypes where external displays are combined with head-worn optics. Beaming displays [1] use an external steerable projector to inject an image into a head-worn optical system where it is magnified and presented to the user. Similarly, Iwai et al. [53] combined head-worn varifocal lenses with an external projector to selectively focus (or blurring) objects in the users' environment. This manipulation is realized by adjusting the focus via the lenses while the projector selectively lights up the environment.

Unlike existing hybrid display approaches that mainly aim to spatially extend the screen volume, our work utilizes standard stationary 2D displays and head-mounted components to realize an off-axis layered display for supporting in-focus contrast.

3 Overview

We present the concept of off-axis layered displays, in which individual layers can be flexibly aligned. In our prototypical implementation, the layers are formed by a combination of a direct-view display, e.g., a stationary desktop screen, and a precisely tracked HMD, that is not necessarily axis-aligned with the stationary 2D display. In contrast to existing approaches, which restrict all layers to be part of the HMD, our system enables focus cues on direct-view displays. As part of our work, we develop a complete end-to-end pipeline for providing focus cues with off-axis layered displays. In the following, we provide an overview of the key components of our approach (illustrated in Figure 3), while subsequent sections provide detailed algorithms and extensions including multi-user support.

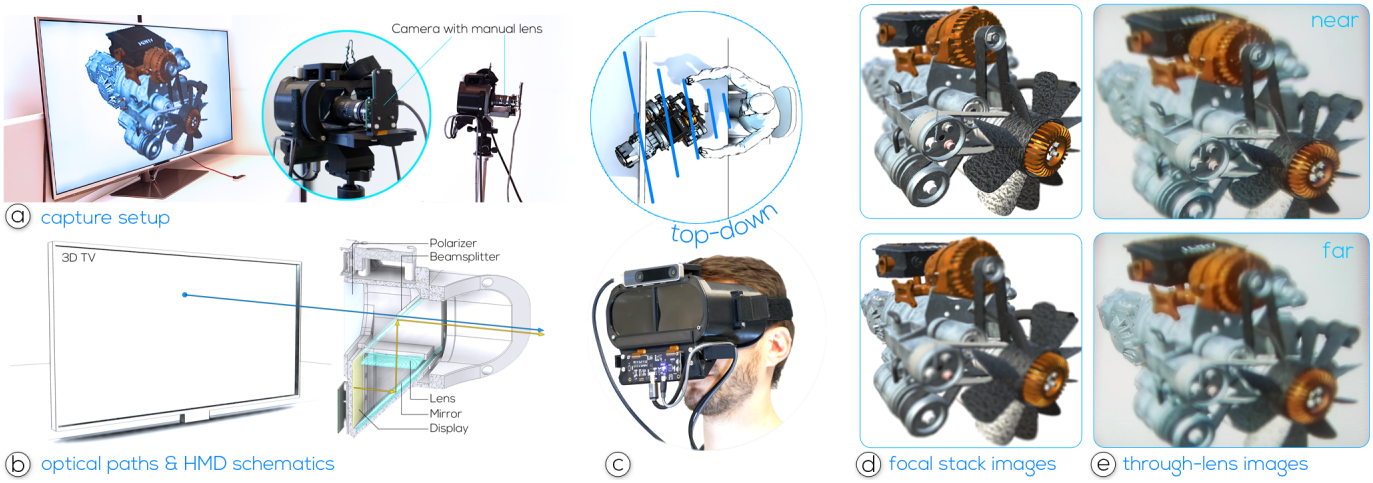

Fig. 4. Additive off-axis layered display. (a) The top-down view of the setup for capturing through-the-lens images shows the spatial relation between the HMD and the direct-view display. (b) Illustration of the optical paths for the additive display and main HMD components. The polarizer has only been introduced to handle 3D displays which use stereoscopic separation based on polarization. A setup using a 3D display based on shuttering can be built without the polarizer. (c) A user wearing the HMD prototype that is illustrated in b). (d) Two images of the focal stack used to represent the 3D scene. (e) Through-the-lens photographs. Near and far-focused images were produced by changing the focus of the camera lens.

Pose tracking. The viewing volume of off-axis layered displays adapts to the head pose of the user. Therefore, we start by estimating the six degrees of freedom (6oF) pose of the HMD relative to the direct-view display (Figure 3a). While several approaches for head tracking are possible, our specific prototype uses an Intel Realsense T265 which has been attached to the HMD. We initialize the system with a known image target that is displayed on the direct-view display.

Focal stack rendering. We compute the user's eye positions by adding a pre-defined offset to the head-pose for each eye. We estimate the offset during a calibration once after the application starts. Subsequently, a focal stack is rendered from each eye and with images parallel to the display plane of the HMD (Figure 3b). Note that the focal stack images facilitate the focal cues, which are encoded within the display. Thus, the quality of the off-axis layered display highly depends on the quality of the focal stack rendering. Therefore, we render a sparse light field that is placed at each estimated eye position and we apply the approach of Ebner et al. [14], as it can generate high-quality defocus blur at high frame rates on the GPU. To further improve runtime performance and quality of the off-axis layered display, we show how to extend off-axis layered displays with eye-tracking which enables rendering and encoding a single focused image only (Section 6).

Focal stack decomposition. In this work, we propose two approaches for computing display patterns. For stationary stereoscopic screens, we decompose each left and right focal stack into two images, resulting in a total of four display patterns per frame, i.e., two per eye (Section 4). Figure 3c shows the two generated patterns for one eye. For monoscopic direct-view displays, we apply a novel optimization scheme realizing a shared display pattern on the direct-view display, as well as two separate images displayed in the HMD, targeting the user's left and right eye respectively (Section 5.1). To improve the perceived quality of the displayed content we furthermore introduce an attenuation layer to the HMD, for which we compute an additional pair of images (Section 5.2).

Post-decomposition warping. The displayed images encode a focal stack using combinations of pixels on both, the HMD and the direct-view display. However, when the user moves, the spatial layer configuration changes in off-axis layered displays, causing the pixel combinations to change which requires updating the decomposition. However, if the decomposition is slower than the user's movements the display pattern cannot be updated accordingly and ghosting artifacts appear. Thus, instead of trying to recompute the patterns at the update rate of the tracker, we realign them according to the current head pose (Figure 3d). Before each frame refresh of the direct-view display, we shift the pixels of the direct-view display pattern so that they realign with those shown in the HMD. Therefore, we store for each pixel in the HMD display pattern the corresponding pixel in the direct-view display pattern (Figure 3d-top). Before each refresh of the direct-view display we project its pattern into the current image plane of the HMD (Figure 3d-middle), where we look-up the stored pixel color of the aligned pattern (Figure 3d-bottom). This step realigns the display patterns at the update rate of the direct-view display.

Display. We have built two display prototypes. First, we combine an optical see-through (OST) HMD with a stereoscopic direct-view display (see Section 4). See Figure 3e for an image captured through the display. In the second prototype, we utilize a standard monoscopic screen. To achieve this, we extend the OST HMD with a low-resolution attenuation layer, which forms, in combination with the OST HMD and the direct-view display, a novel type of hybrid attenuated-additive layered display (Section 5.1).

In the following, we explain of the core components for each of the implemented prototypes: stereoscopic and monoscopic.

4 Stereoscopic Direct-View Display Layer

Our first HMD prototype uses one LCD panel per eye (Sharp, 120Hz, 1440^2 resolution), for the color display. The light coming from the LCD is reflected by a mirror and refracted by a lens (5 cm focal length) which creates a virtual image of the LCD at a distance of about 30cm. The light is combined with light of the environment via a beam splitter. Note that polarizers are used to selectively filter the light coming from a passive stereo display to provide correct views for each eye. Refer to Figure 4b for an illustration of the image forming process.

The display patterns of a 3D scene are computed such that when viewed through the apparatus, the scene is reproduced with correct depth of field. Since we are using an OST HMD, the perceived image, subsequently called the retinal image, is formed by the additive blending of the display patterns (Figure 4). Several approaches to pattern computation have been proposed for additive layered displays [38, 42]. We adapt the core principles of these techniques and generalize them for arbitrary display configurations by introducing the 3D geometry of the direct-view display. Since we are using a stereoscopic display, we can decompose the 3D scene into display patterns for each eye independently. For simplicity, we describe our approach for a single eye, single color channel, and two planar displays. However, we would like to emphasize that the approach is suitable for arbitrary numbers and shapes of displays.

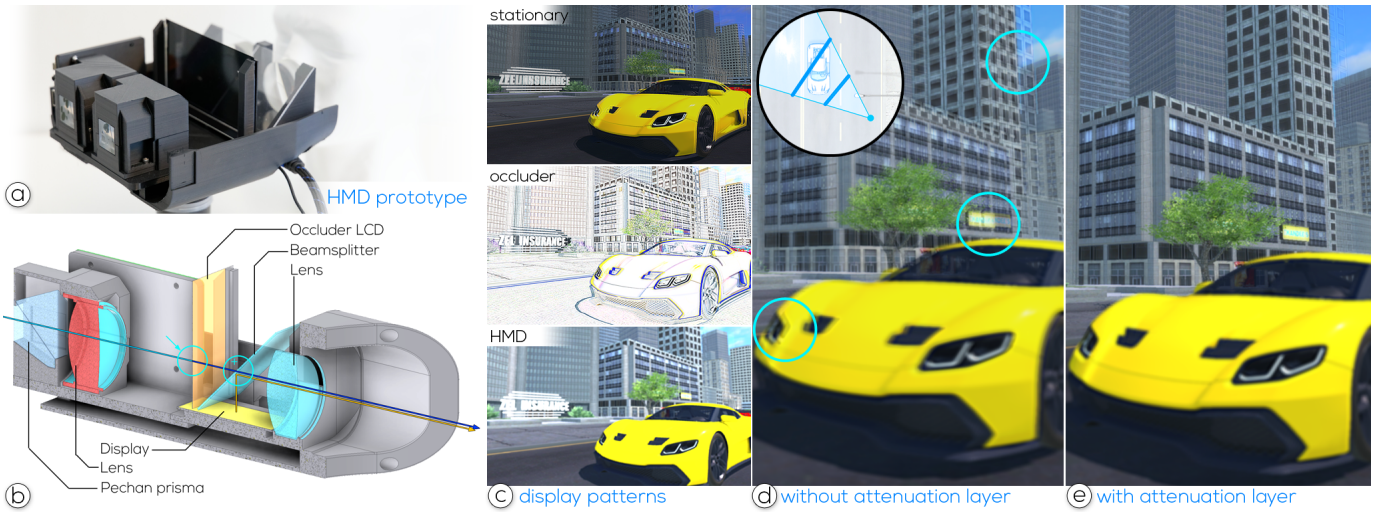

Fig. 5. Off-axis layered displays with attenuation layer. (a) Photograph of the HMD prototype with attenuation layer. (b) A light ray from the direct-view display entering the HMD passes through the attenuation layer before it is combined with the light ray from the HMD screen. (c) The display patterns shown on the different screens. (d, e) Examples of the joint stereoscopic decomposition with and without attenuation layer. The inset on top shows the setup of the rendered scene. (d) The result without the attenuation layer shows ghosting artifacts, which have been highlighted with circles (e) The attenuation layer is able to filter out the contribution meant for the other eye, which reduces the ghosting artifacts.

4.1 Off-axis image formation

As suggested in prior work [38, 56], we model the human eye with an ideal thin lens and a circular aperture of $a = 4$ mm, which corresponds to the average pupil diameter of the human eye. We approximate the retina as a planar image sensor with a resolution of $X \times Y$ pixels. The distance from the pupil to the retina is set to $d_e = 17$ mm, and we assume the retinal circle of confusion (CoC) of a 3D point using a disk with a diameter

$$c(d_o, d_f) = a \cdot d_e \frac{|d_o - d_f|}{d_o \cdot d_f}, \tag{1}$$

where d_o is the depth of the point, and d_f is the current focus distance.

Given the number of pixels in the HMD and the monitor by $U \times V$ and $S \times T$, respectively, we define $h_1 \in \mathbb{R}^{UV}$ and $h_2 \in \mathbb{R}^{ST}$ to represent the display color values as column vectors, where h_1 refers to the pixel colors of the HMD, and h_2 refers to the pixel color values of the direct-view display. For a given focal distance d_f, we describe the resulting retinal image $Z_f \in \mathbb{R}^{X \times Y}$ as a column vector $z_f \in \mathbb{R}^{XY}$. To represent N retinal images, each with a focal distance d_f, $f \in [1; N]$, we concatenate each retinal image z_f into a column vector z:

$$\underbrace{\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{bmatrix}}_{z} = \underbrace{\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ \vdots & \vdots \\ A_{N,1} & A_{N,2} \end{bmatrix}}_{A} \cdot \underbrace{\begin{bmatrix} h_1 \\ h_2 \end{bmatrix}}_{h} \tag{2}$$

The matrices $A_{f,1} \in \mathbb{R}^{XY \times UV}$ and $A_{f,2} \in \mathbb{R}^{XY \times ST}$ define the projections that map the corresponding display pixels onto the retina image and spread the pixels according to the CoC as defined in Equation 1, such that $z_f = A_{f,1} h_1 + A_{f,2} h_2$.

4.2 Solving for off-axis display configurations

To encode multiple retinal images, we use a focal stack with N images r_f as a reference. Each image in the focal stack is focused on a different distance d_f in between the display planes. To compute a display pattern, we minimize the Euclidean distance between each r_f and the corresponding retinal image z_f:

$$\underset{h}{\operatorname{argmin}} \sum_{f=1}^{N} ||r_f - z_f||, \quad \text{s.t. } h \in [0; 1]^{UV+ST} \tag{3}$$

Equation 3 is solved using iterative SART [3]. In each iteration n, the current decomposition is updated using Equation 4.

$$h_k^{(n)} = h_k^{(n-1)} + \frac{\beta}{M_k} \sum_{j=1}^{J} \frac{A_{j,k}(r_j - (Ah^{(n-1)})_k)}{M_j}, \quad \text{where} \quad \underset{k \in \{1,2\}}{M_k} = \sum_{j=1}^{J} A_{j,k}, \quad M_j = \sum_{k=1}^{2} A_{j,k}, \tag{4}$$

and β is a relaxation parameter. Initially, the decompositions are set to zero, i.e., $h_k^{(0)} = 0$. After each iteration, the values of h are clamped to $[0; 1]$. Equation 4 is implemented on the GPU using rasterization and compute shaders. Each iteration is split into the following two steps.

Retinal image computation. To derive the retinal image, z_f when focusing on a distance f, we compute the retinal image by blending the two display patterns additively. To support arbitrary spatial display configurations for our off-axis layered display, we approximate each display surface using a 3D mesh, which is textured with the corresponding display pattern. The retinal image is rendered by projecting the textured surface onto the retina, i.e., into camera space. Each pixel is spread using a 2D scatter operation with a kernel size corresponding to the pixel's CoC.

Display pattern update. The display images are updated in each iteration by adding the error at the time of the current iteration. The error of the current decomposition is computed by subtracting the retinal image from the corresponding references image in the focal stack. The error is then projected to the display layers and distributed according to its CoC by applying a 2D convolution filter with a kernel size corresponding to Equation 1.

5 Monoscopic Direct-View Display

While stereoscopic screens allow displaying the image patterns for each eye separately, most direct-view displays are monoscopic. Thus, we modify our initial approach based on the concept of color coding for stereoscopic image separation, as used in anaglyph 3D images [18]. This goal can be achieved and jointly decompose the left and right focal stacks into three, instead of four, images, where one image is shown to both eyes simultaneously on the monoscopic display. However, since the image formation is additive, each eye receives a portion of the pattern that is meant for the other eye (see Figure 5d). To prevent such artifacts, we also extend the HMD with an attenuation layer, which can filter out the wrong pattern contributions.

5.1 Joint stereoscopic decomposition

As before, the input to our system is a focal stack for each eye. However, to jointly compute all three display patterns we need to reformulate our model. Therefore, we introduce the solution vectors h_1^{L} for the left eye, and, h_1^{R} for the right eye. The pixel vector of the monitor remains h_2, so that the model of the image formation process adapts to Equation 5. Equation 4 can be used for solving the new model. However, its implementation needs to be adapted to project all error values to all displays.

$$\begin{bmatrix} z_1^{\mathrm{L}} \\ \vdots \\ z_N^{\mathrm{L}} \\ z_1^{\mathrm{R}} \\ \vdots \\ z_N^{\mathrm{R}} \end{bmatrix} = \begin{bmatrix} A_{1,1}^{\mathrm{L}} & 0 & A_{1,2}^{\mathrm{L}} \\ \vdots & \vdots & \vdots \\ A_{N,1}^{\mathrm{L}} & 0 & A_{N,2}^{\mathrm{L}} \\ 0 & A_{1,1}^{\mathrm{R}} & A_{1,2}^{\mathrm{R}} \\ \vdots & \vdots & \vdots \\ 0 & A_{N,1}^{\mathrm{R}} & A_{N,2}^{\mathrm{R}} \end{bmatrix} \cdot \begin{bmatrix} h_1^{\mathrm{L}} \\ h_1^{\mathrm{R}} \\ h_2 \end{bmatrix} \tag{5}$$

5.2 Attenuation support

Several approaches for implementing pixel-wise attenuation on an OST HMD have been proposed in the past [29, 30]. Following the approach of Kiyokawa et al. [29], we introduce an additional LC layer and a relay lens system to focus the image. For the lenses in the relay system, we use two achromatic doublets (focal length 7.5 cm each). The principal planes of the lenses are spaced two times their focal length, to provide an unmagnified view of the real world. To obtain occlusion capability, we use an LCD (Sharp, 3840×2160 resolution) without a backlight which is located between the lenses. An additional LCD is used together with a beam splitter to augment the light passing through the occluder. The setup is depicted in Figure 5b.

To integrate this setup in the decomposition, we define two occlusion layers o^{L} and o^{R}, for the left and right eye, respectively, and we assume that light rays coming from the monitor are attenuated by the occlusion layers in a multiplicative manner. Thus, we model the attenuation image formation process as

$$\begin{aligned} z_f^{\mathrm{L}} &= A_{f,1}^{\mathrm{L}} h_1^{\mathrm{L}} + \lambda \cdot (B_f^{\mathrm{L}} o^{\mathrm{L}}) \odot (A_{f,2}^{\mathrm{L}} h_2), \\ z_f^{\mathrm{R}} &= A_{f,1}^{\mathrm{R}} h_1^{\mathrm{R}} + \lambda \cdot (B_f^{\mathrm{R}} o^{\mathrm{R}}) \odot (A_{f,2}^{\mathrm{R}} h_2), \end{aligned} \tag{6}$$

where $\odot$ is the Hadamard product, B_f^{L} and B_f^{R} are the projection matrices for the occlusion layers, and $\lambda \in [0;1]$ is a damping factor that denotes the maximum amount of light passing through the attenuation layer. Note that, other than the image formation being multiplicative, this model makes no assumption about the underlying display technology used for the attenuation layer, and therefore, it is suitable for both LC and LCoS displays.

As the image formation model in Equation 6 is no longer linear, the decomposition cannot be directly obtained with SART. Therefore, we propose a new optimization approach. For each eye, we split the image formation model into two subsystems, $s_{f,1} = A_{f,1} h_1$ and $s_{f,2} = \lambda \cdot B_f o \cdot A_{f,2} h_2$. For a given decomposition, we compute the error regarding a focal stack image by subtracting the retinal image from the focal stack image. Since the subsystems contribute equally to the retinal image, each subsystem is updated using half of the error. For subsystem $s_{f,1}^{\mathrm{L}}$ and $s_{f,1}^{\mathrm{R}}$, we can directly back-project ε_f to the HMD and update the decomposition. This method is equivalent to using Equation 4 for the update.

6 Vergence-driven optimization

A direct SART update is not possible because in the subsystems $s_{f,2}^{\mathrm{L}}$ and $s_{f,2}^{\mathrm{R}}$, the decompositions of the direct-view display and the attenuation layers are coupled multiplicatively. Thus, we rearrange the subsystems $s_{f,2}^{\mathrm{L}}$ and $s_{f,2}^{\mathrm{R}}$ to a single matrix $S_{f,2} \in \mathbb{R}^{XY \times 2XY}$. Further, we define the matrix corresponding to half the reconstruction error as E_f. For a given focal stack image, the updated subsystem matrix $S_{f,2}$ is computed as:

$$S_{f,2} + E_f = \left(\underbrace{\begin{bmatrix} B_f^{\mathrm{L}} & 0 \\ 0 & B_f^{\mathrm{R}} \end{bmatrix}}_{B_f} \cdot \lambda \underbrace{\begin{bmatrix} o^{\mathrm{L}} \\ o^{\mathrm{R}} \end{bmatrix}}_{o} \right) \otimes \left(A_{f,2} h_2 \right), \tag{7}$$

where $\otimes$ denotes the outer product. We solve for o and h_2 using non-negative matrix factorization. More specifically, we define $v = B_f \lambda \cdot o$ and $w = A_{f,2} h_2$ and use the multiplicative update rule of Lee and Seung [33] to update the decompositions as follows:

$$\begin{aligned} h_2^{(m)} &= h_2^{(m-1)} \odot \frac{\sum_{f=1}^{N} A_{f,2}^T v^{T\,(m-1)} \left(S_{f,2} + E_f \right)}{\sum_{f=1}^{N} A_{f,2}^T v^{T\,(m-1)} v^{(m-1)} w^{(m-1)}}, \\ o^{(m)} &= o^{(m-1)} \odot \frac{\sum_{f=1}^{N} B_f^T \left(S_{f,2} + E_f \right) w^{T\,(m)}}{\sum_{f=1}^{N} B_f^T v^{(m-1)} w^{(m)} w^{T\,(m)}}, \end{aligned} \tag{8}$$

where o is initially set to fully transparent, and h_2 is initialized with zeros. In our implementation, matrix multiplication with A and A^T is

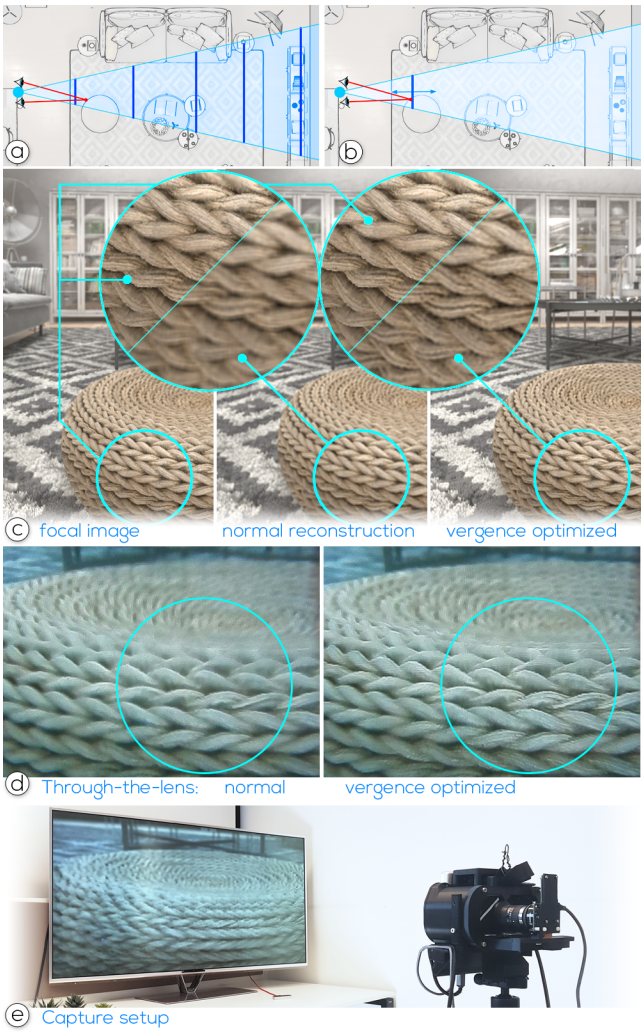

Fig. 6. Vergence-driven optimization. (a) Overview of the example scene. Red lines indicate measured vergence points; blue lines indicate focal stack image planes. (b) The focal image is rendered at the measured focal distance. (c) Comparison between (left) ground-truth, (middle) focal stack based decomposition, and (right) vergence-driven optimization. Vergence-driven optimization is able to provide more contrast. (d) Screenshots captured through the display. (e) The setup used to capture the images in d).

again carried out as described in the section above. In practice, we use a single iteration of Equation 8, before we recompute E_f and start over until all decompositions are stable.

To focus at several distances, off-axis layered displays have to encode several focused images why they commonly suffer from reduced contrast and an increased computational effort. To mitigate these issues, we introduce *vergence-driven optimization* to layered displays, which considers only a single focused image to be encoded in the layered display. To render the image that corresponds to the user's focus distance, we measure the vergence distance with an eye tracker. Vergence and accommodation are coupled and thus, we assume that the measured vergence distance corresponds to the focus distance of the user.

Since we assume a known scene in between the HMD and the direct-view display, we can compute a per-pixel depth map, and use it to infer the current vergence by intersecting the gaze location with the depth map [43]. This approach is known to be more precise, as it is not affected by the inaccuracies of measuring vergence distance using the gaze angle [12].

Once the focus distance of the user is known, it is not necessary to render and decompose the entire focal stack. Instead, we only render a single image with the focal distance aligned with the user's vergence distance (see Figure 6a versus Figure 6b for an illustration). Since we only need to compute the display pattern from a single image, vergence-driven optimization reduces the computation time for the decomposition, and the display can achieve better contrast when focusing on an object at this particular distance (see Figure 6c-d).

7 Multiple Users

Similar to using an attenuation layer to filter out the contribution of one view in a two-view display pattern, we can utilize the attenuation layer to block several views, which enables supporting multiple users. In such a configuration, each user is wearing a dedicated HMD, whereas the direct-view display is shared by all users. Thus, the display pattern that is shown on the direct-view display needs to incorporate one image pattern for each user, while the attenuation layer of each individual user needs to filter out the contribution that is meant for all other users. For example, with two users and a monoscopic direct-view display, three views need to be filtered out by each attenuation layer, i.e., two views that correspond to the other user and one view representing the user's other eye.

To enable multi-user support, we adjust the decomposition to take the views of all users into account. To this end, we use a continuous index for the image formation process of each view i:

$$z_f^i = A_{f,1}^i h_1^i + \lambda \cdot (b^i o^i) \odot (A_{f,2}^i h_2). \tag{9}$$

For U users, in a system with a monoscopic direct-view display, the maximum number of views i equals $2U$. A stereoscopic screen reduces the number of views used in the decomposition algorithm to U, since the decompositions for the two eyes can be computed separately.

8 Evaluation

In this section, we present an evaluation of the image quality and computational demand for individual components. Since capturing the physical image through the display properly using a digital camera is difficult, we follow the evaluation approach used in recent literature [14, 38, 42] and focus on the algorithmic part of our pipeline, i.e., we compare the results of our decomposition method with the input focal stack images as ground truth. This approach lets us use state-of-the-art image metrics (we used PSNR, SSIM and LPIPS) that allow a meaningful comparison to previous work.

Throughout the evaluation, the system was driven by a standard desktop workstation, using an AMD Ryzen 9 3900X CPU at 4GHz, 64GB RAM, 2× NVIDIA 3090 Ti GPU, and Windows 10 as operating system. For an application, such as the racing game depicted in Figure 7, (about 2 dpt volume), rendering times for a single eye (1024^2 pixels) were around 22 ms on average using the prototype without attenuation layer and utilizing vergence-driven optimization. To prevent falling below real-time frame rates, we used one GPU per eye.

Fig. 7. A photograph of a user interacting with a two-layer display, where the first layer is displayed in a self-made head-worn prototype, and the second layer is a large passive-stereo TV set in the background.

8.1 Image quality of the layered display

To quantitatively evaluate our system, we simulate an off-axis layered display with and without attenuation layer to explore the perceived image quality, and the placement and the resolution of the attenuation layer. We configure the simulator with a direct-view display at 2 m, an HMD focus plane at 0.3 m, and the attenuation layer at a distance of 5 m to the user's eyes. We simulate the perceived image when looking through the display by using an aperture that is set to a diameter of 4 mm. The image patterns are computed using the joint stereoscopic decomposition so that a monoscopic direct-view display can be used. We compare this configuration with a simulation of an off-axis layered display without the attenuation layer in three different scenes, depicted in Figure 9. For each scene, we generate a focal stack with 11 images, placed equidistant between the virtual image plane of the HMD, i.e., placed at 30 cm, and the direct-view display. Each image in the focal stack has a resolution of 2048^2, while the attenuation layer has a resolution of 1024^2 pixels.

Table 1 shows the measurements of *peak signal-to-noise ratio* (PSNR), *structured similarity image metric* (SSIM), and the *learned perceptual image patch similarity* (LPIPS) [58] for near and far focus distances. The results show an improvement when using the attenuation layer for all scenes and both focus distances. Furthermore, it can be seen that the quality at the far focus distance is always lower compared to the image quality at a near focus distance. This limitations has been expected, as the attenuation layer cannot completely filter the shared image into two fully separated images. As Figure 5d-e shows, the attenuation mitigates ghosting, but does not completely remove it. Therefore, when focusing to the front, the shared image pattern in the background is blurred and slight ghosting artifacts in the shared image have a lower impact on the quality of the perceived image.

8.2 Impact of the attenuation layer

To evaluate the influence of the attenuation layer, we measure the perceived contrast in several configuration with varying distances of the attenuation layer. For measuring the contrast, we render a plane of sine-gratings from 1≈20 cpd in the dioptric center between the direct-view display and the image plane of the HMD. Since off-axis displays allow interactively altering the layer setup, we assess them by measuring the contrast for several distances. Thus, we gradually increase the distance between the direct-view display and the HMD, while computing the mean contrast in the perceived image when the user focuses on the plane with the sine gratings. Focal stack images are rendered at a resolution of 20 CPD and placed at every 0.2 dpt.

Figure 8 show the measured contrast in an off-axis display with an attenuation layer at various distances for difference volumes, defined by the distance between the HMD and the direct-view display. To avoid the impact of the ghosting artifacts on monoscopic displays, we have configured the simulator using a stereoscopic direct-view display, which allows separating the decomposition of the left and right eye. The results indicate that an attenuation layer set at 40 cm (purple

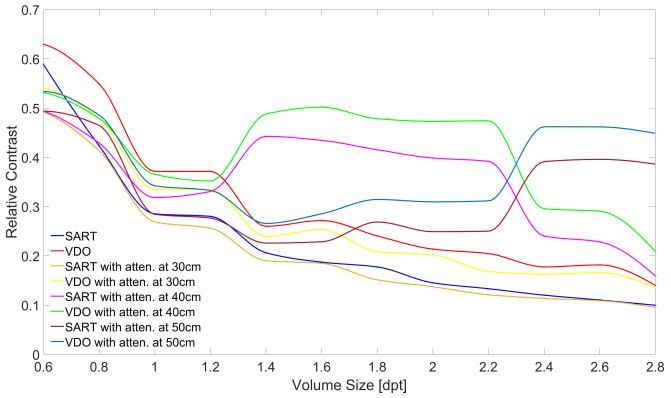

Fig. 8. Comparison of perceived contrast when using vergence-driven optimization (VDO) and the conventional optimization using a focal stack. Results have been measured while focusing in the volume center

line in Figure 8) can improve the perceived contrast significantly for medium-sized volumes of up to 2.2 dpt. Note that the vergence-driven optimization (VDO) in combination with an attenuation layer, placed at 40 cm, indicated by the green line in Figure 8, achieves the highest overall contrast for volumes of up to 2.2 dpt. Placing the attenuation layer at 30 cm shows no benefit compared to not using an attenuation layer. An attenuation layer placed further, in our evaluation, at 50 cm, supports larger volumes, but suffers from lower contrast values in smaller volumes.

To assess the impact of the resolution of the attenuation layer to the perceived image quality we simulate three different resolutions of attenuation layers. Table 2, shows the measured quality metrics for three focus distances. The results indicate that quality increases only slightly with higher resolutions of the attenuation layer. It must be noted that our simulation does not consider diffraction artifacts that are known to affect the image quality of attenuative layered displays [21]. While lower resolutions will not be affected, we expect lower quality results for resolutions higher than 1024^2. However, since increasing the resolution only slightly improves the image quality, we recommend configurations that stay above the diffraction limit.

8.3 Vergence-driven optimization

The effectiveness of the vergence-driven optimization can be demonstrated by measuring its impact on the perceived contrast and on the time required to compute the display pattern. We measure the perceived contrast as described for the evaluation of the attenuation layer. Figure 8 shows the results obtained with and without vergence-driven optimization. We observe that vergence-driven optimization leads to the higher contrast throughout the entire range of tested volumes. The quality improvement can be explained by the number of images that need to be encoded by the display. Vergence-driven optimization only needs to encode a single image, while a conventional decomposition considers the entire focal stack.

Compared to computing the display pattern from a dense focal stack with many images, vergence-driven optimization can also lead to faster update rates. Table 3 shows the runtimes for both approaches for three different resolutions. The results show that vergence-driven optimization can reduce the runtime. The improved performance allows us to increase the resolution of the reference focal image and the number of iterations spent on optimizing the decomposition. To get an intuition on how much the quality can be improved, we perform an equal-time comparison. It measures the quality of the display when iterating the decomposition of higher quality focal stack images more often, with respect to the time we gain from using vergence-driven optimization. In our test, we use a resolution of $1,024^2$ for images in the focal stack, and a resolution of $2,048^2$ for the focal image used in the vergence-driven optimization. The comparison is conducted at 30 fps using the three scenes (Figure 9). The results (Table 4) show an increase in quality for all scenes and across all quality metrics when using vergence-driven optimization.

Fig. 9. Scenes used in evaluation reported in Tables 1, 2, 4 and 5.

Table 1. Simulation of a monoscopic direct-view display and an HMD with and without an attenuation layer at two focus distances.

	Setup	No Attenuation		With Attenuation	
	Focus	Near	Far	Near	Far
PSNR dB	Scene A	32.72	28.75	35.10	31.05
	Scene B	33.27	24.48	33.36	26.51
	Scene C	29.25	22.43	31.46	25.01
SSIM	Scene A	0.9719	0.8776	0.9780	0.9118
	Scene B	0.9832	0.9676	0.9833	0.9899
	Scene C	0.9581	0.8024	0.9674	0.8627
LPIPS	Scene A	0.0980	0.2470	0.0820	0.1950
	Scene B	0.0840	0.2110	0.0700	0.1910
	Scene C	0.1260	0.3010	0.1180	0.2330

Table 2. Quality of perceived images over varying resolutions of the attenuation layer for three focus distances using scene A in Figure 9.

		512^2	$1,024^2$	$2,048^2$
PSNR dB	Near focus	36.88	37.18	37.26
	Center focus	34.42	34.95	35.09
	Far focus	30.68	31.43	31.78
SSIM	Near focus	0.9841	0.9859	0.9863
	Center focus	0.9436	0.9508	0.9525
	Far focus	0.9108	0.9281	0.9363
LPIPS	Near focus	0.0560	0.0490	0.0470
	Center focus	0.1490	0.1310	0.1270
	Far focus	0.1900	0.1500	0.1310

Table 3. Compute times for a single SART iteration with and without vergence-driven optimization (VDO) for three resolutions and three different volume sizes.

		512^2	$1,024^2$	$2,048^2$
3dpt	SART	22.67ms	71.59ms	256.78ms
	SART VDO	1.75ms	6.73ms	28.44ms
2dpt	SART	10.13ms	40.86ms	113.17ms
	SART VDO	2.85ms	5.05ms	20.83ms
1dpt	SART	3.26ms	14.96ms	46.48ms
	SART VDO	0.95ms	3.54ms	16.24ms

Table 4. Equal time comparison between conventional decomposition and vergence-driven optimization.

		Conventional SART	Vergence-driven optimization
PSNR dB	Scene A	34.42	38.06
	Scene B	32.22	38.87
	Scene C	30.10	33.94
SSIM	Scene A	0.9424	0.9762
	Scene B	0.9784	0.9947
	Scene C	0.9723	0.9907
LPIPS	Scene A	0.112	0.065
	Scene B	0.092	0.052
	Scene C	0.1350	0.0710

Table 5. Perceived image quality in a two-user setup. The comparison is carried out for two distances between the users and two focus distances.

	User distance	0.5 m		1.0 m	
	Focus	Near	Far	Near	Far
PSNR dB	Scene A	44.91	33.41	44.84	33.23
	Scene B	33.02	27.30	33.07	27.57
	Scene C	29.94	25.31	30.50	25.72
SSIM	Scene A	0.9947	0.9321	0.9946	0.9282
	Scene B	0.9909	0.9423	0.9910	0.9437
	Scene C	0.9740	0.8895	0.9762	0.8971
LPIPS	Scene A	0.012	0.1485	0.0125	0.1575
	Scene B	0.0125	0.1575	0.0210	0.1265
	Scene C	0.0545	0.1890	0.0580	0.1720

8.4 Multiple users

To evaluate the capability for handling multiple users, we simulate a setup in which two users are simultaneously looking at the direct-view display. The perceived image quality is simulated with users positioned 0.5 m and 1 m apart. Table 5 shows the measured PSNR, SSIM, and LPIPS values. The results indicate that the distance between users has only a small impact on the quality of the perceived image. However, when analyzing the impact of individual focus distances on the result, it is evident that focusing at the front leads to a better image quality. This observation is similar to the results seen in the evaluation of the attenuation layer. When focusing to an object closer to the user, the shared image pattern that is shown on the direct-view display is out of focus and, therefore, will consist of defocus blur, which diminishes the impact of artifacts.

9 Conclusion and Future Work

We have introduced off-axis layered displays. Our results indicate that they can increase the perceived contrast when focusing at virtual objects between the user and a direct-view display. While stereoscopic displays have been used for many years, off-axis layered display enable perceptual realism outside the display plane of direct-view displays. While this can increase perceptual comfort, it is an important requirement for exploring realistic 3D scene representations, such as those provided by light-field [40] and neural radiance field representations [39].

There are several aspects of our work that we did not specifically focus on and as such, that offer the potential for further exploration beyond the scope of this paper. For example, we did only apply a manual color calibration to radiometrically align the screens of off-axis layered displays. Improving the color calibration will improve the quality of perceived images, but requires more work to properly align the colors. Furthermore, we did not evaluate our approach with human participants, but instead, focused on a technical evaluation as it is common for such an approach. However, a user evaluation may provide additional insights in the usability and perceptual performance of the display, especially in volumes of varying size. For example, we only tested the impact of eye-tracking data in a simulator, and by using a known focus distance for capturing images through the display. As our test assume perfect eye-tracking, we might see different results when testing with humans. However, as we already see a trend towards high-quality eye-tracking [4, 43], we only expect slight variations from our current results.

Besides the current limitations, we also envision several directions for further improvement from developing additional components. For example, perceptual-based rendering, in particular foveated rendering, can be used to further improve the quality of the perceived images. Similarly, tracking the gaze of multiple users of the system would allow optimizing the shared image pattern shown on the direct-view display. While this would improve the performance, it can also increase the contrast when both users look at different areas of the direct-view display since the decomposition can be computed locally for each user.

Since off-axis layered displays support VR and AR applications, we are also interested in exploring its perceptual benefit to applications across realities [15,52]. In addition, we see potential in investigating different display types for direct-view displays. The possible options range from projectors to handheld displays. Projectors support stereoscopic viewing but also offer a larger screen area. In particular, extending existing CAVE installations with focus cues, as illustrated in Figure 2 is an interesting practical application. Furthermore, we believe that smaller handheld screens have a lot of potential. While their screen size is smaller, their resolution and relative proximity should compensate for this, albeit at the cost of a smaller viewing volume. Still, it would fill the research gap next to approaches such as Multifi and Trackcap that expand screen space [17, 41] while allowing for exploring novel interactions with 3D content.

Acknowledgments

This work was partially sponsored by Snap Inc., the Hong Kong UGC Early Career Scheme Fund (27212822), and the Marsden Fund Council from Government funding (grant no. MFP-UOO2124).

References

[1] K. Aksit, Y. Itoh, and T. Kaminokado. Beaming displays: towards display-less augmented reality near-eye displays. In *AI and Optical Data Sciences III*, vol. 12019, pp. 34–37. SPIE, 2022.

[2] K. Akşit, W. Lopes, J. Kim, P. Shirley, and D. Luebke. Near-eye varifocal augmented reality display using see-through screens. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017.

[3] A. H. Andersen and A. C. Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984.

[4] A. N. Angelopoulos, J. N. Martel, A. P. Kohli, J. Conradt, and G. Wetzstein. Event-based near-eye gaze tracking beyond 10,000 hz. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 27(5):2577–2586, 2021.

[5] P. Budhiraja, R. Sodhi, B. Jones, K. Karsch, B. Bailey, and D. Forsyth. Where's my drink? enabling peripheral real world interactions while using hmds. *arXiv preprint arXiv:1502.04744*, 2015.

[6] O. Cakmakci and J. Rolland. Head-worn displays: a review. *Journal of display technology*, 2(3):199–216, 2006.

[7] P. Chakravarthula, Y. Peng, J. Kollin, H. Fuchs, and F. Heide. Wirtinger holography for near-eye displays. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[8] C. Chang, K. Bang, G. Wetzstein, B. Lee, and L. Gao. Toward the next-generation vr/ar optics: a review of holographic near-eye displays from a human-centric perspective. *Optica*, 7(11):1563–1578, 2020.

[9] S. Choi, M. Gopakumar, Y. Peng, J. Kim, and G. Wetzstein. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021.

[10] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pp. 69–117. Elsevier, 1995.

[11] N. A. Dodgson. Autostereoscopic 3d displays. *Computer*, 38(8):31–36, 2005.

[12] D. Dunn. Required accuracy of gaze tracking for varifocal displays. In *Proc. IEEE Virtual Reality (VR)*, pp. 1838–1842. IEEE, 2019.

[13] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Akşit, P. Didyk, K. Myszkowski, D. Luebke, and H. Fuchs. Wide field of view varifocal near-eye display using see-through deformable membrane mirrors. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(4):1322–1331, 2017.

[14] C. Ebner, S. Mori, P. Mohr, Y. Peng, D. Schmalstieg, G. Wetzstein, and D. Kalkofen. Video see-through mixed reality with focus cues. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 28(5):2256–2266, 2022.

[15] B. Fröhler, C. Anthes, F. Pointecker, J. Friedl, D. Schwajda, A. Riegler, S. Tripathi, C. Holzmann, M. Brunner, H. Jodlbauer, H. Jetter, and C. Heinzl. A Survey on Cross-Virtuality Analytics. *Computer Graphics Forum*, 2022. doi: 10.1111/cgf.14447

[16] M. Gopakumar, J. Kim, S. Choi, Y. Peng, and G. Wetzstein. Unfiltered holography: optimizing high diffraction orders without optical filtering for compact holographic displays. *Optics Letters*, 46(23):5822–5825, 2021.

[17] J. Grubert, M. Heinisch, A. Quigley, and D. Schmalstieg. Multifi: Multi fidelity interaction with displays on and around the body. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, p. 3933–3942, 2015.

[18] R. R. Hainich and O. Bimber. *Displays: fundamentals & applications*. AK Peters/CRC Press, 2016.

[19] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of vision*, 8(3):33–33, 2008.

[20] J. Hong, Y. Kim, H.-J. Choi, J. Hahn, J.-H. Park, H. Kim, S.-W. Min, N. Chen, and B. Lee. Three-dimensional display technologies of recent interest: principles, status, and issues. *Applied optics*, 50(34):H87–H115, 2011.

[21] F. Huang, K. Chen, and G. Wetzstein. The Light Field Stereoscope: Immersive Computer Graphics via Factored Near-Eye Light Field Displays with Focus Cues. *ACM Transactions on Graphics (TOG)*, (4), 2015.

[22] Y. Itoh, T. Langlotz, J. Sutton, and A. Plopski. Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays. *ACM Computing Surveys*, 54(6), July 2021.

[23] B. Javidi, A. Carnicer, A. Anand, G. Barbastathis, W. Chen, P. Ferraro, J. Goodman, R. Horisaki, K. Khare, M. Kujawinska, et al. Roadmap on digital holography. *Optics Express*, 29(22):35078–35118, 2021.

[24] B. R. Jones, H. Benko, E. Ofek, and A. D. Wilson. Illumiroom: Peripheral projected illusions for interactive experiences. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, p. 869–878, 2013.

[25] D. Kim, S. Lee, S. Moon, J. Cho, Y. Jo, and B. Lee. Hybrid multi-layer displays providing accommodation cues. *Optics Express*, 26(13):17170–17184, Jun 2018. doi: 10.1364/OE.26.017170

[26] J. Kim, Y. Jeong, M. Stengel, K. Aksit, R. A. Albert, B. Boudaoud, T. Greer, J. Kim, W. Lopes, Z. Majercik, et al. Foveated ar: dynamically-foveated augmented reality display. *ACM Transactions on Graphics (TOG)*, 38(4):99–1, 2019.

[27] S.-K. Kim, K.-H. Yoon, S. K. Yoon, and H. Ju. Parallax barrier engineering for image quality improvement in an autostereoscopic 3d display. *Optics express*, 23(10):13230–13244, 2015.

[28] Y. Kitamura, T. Konishi, S. Yamamoto, and F. Kishino. Interactive stereoscopic display for three or more users. In *Proc. Conference on Computer Graphics and Interactive Techniques (Siggraph)*, pp. 231–240, 2001.

[29] K. Kiyokawa, M. Billinghurst, B. Campbell, and E. Woods. An occlusion capable optical see-through head mount display for supporting co-located collaboration. In *Proc. International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 133–141. IEEE, 2003.

[30] B. Krajancich, N. Padmanaban, and G. Wetzstein. Factored occlusion: Single spatial light modulator occlusion-capable optical see-through augmented reality display. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26(5):1871–1879, 2020.

[31] G. Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(7):1912–1931, 2016.

[32] D. Lanman and D. Luebke. Near-eye light field displays. *ACM Transactions on Graphics (TOG)*, 32(6), nov 2013.

[33] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

[34] S. Lee, J. Cho, B. Lee, Y. Jo, C. Jang, D. Kim, and B. Lee. Foveated retinal optimization for see-through near-eye multi-layer displays. *IEEE Access*, 6:2170–2180, 2017.

[35] S. Lee, Y. Jo, D. Yoo, J. Cho, D. Lee, and B. Lee. Tomographic near-eye displays. *Nature communications*, 10(1):1–10, 2019.

[36] A. W. Lohmann, R. G. Dorsch, D. Mendlovic, Z. Zalevsky, and C. Ferreira. Space–bandwidth product of optical signals and systems. *Journal of the Optical Society of America A (JOSA A)*, 13(3):470–473, 1996.

[37] M. McGill, D. Boland, R. Murray-Smith, and S. Brewster. A dose of reality: Overcoming usability challenges in vr head-mounted displays. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, CHI '15, p. 2143–2152, 2015.

[38] O. Mercier, Y. Sulai, K. Mackenzie, M. Zannoli, J. Hillis, D. Nowrouzezahrai, and D. Lanman. Fast gaze-contingent optimal decompositions for multifocal displays. *ACM Transactions on Graphics (TOG)*, 36(6):237, 2017.

[39] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):29:1–29:14, 2019.

[40] P. Mohr, S. Mori, T. Langlotz, B. H. Thomas, D. Schmalstieg, and D. Kalkofen. Mixed reality light fields for interactive remote assistance. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1–12, 2020.

[41] P. Mohr, M. Tatzgern, T. Langlotz, A. Lang, D. Schmalstieg, and D. Kalkofen. Trackcap: Enabling smartphones for 3d interaction on mobile head-mounted displays. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, p. 1–11, 2019.

[42] R. Narain, R. A. Albert, A. Bulbul, G. J. Ward, M. S. Banks, and J. F. O'Brien. Optimal presentation of imagery with focus cues on multi-plane displays. *ACM Transactions on Graphics (TOG)*, 34(4):59, 2015.

[43] N. Padmanaban, R. Konrad, and G. Wetzstein. Autofocals: Evaluating gaze-contingent eyeglasses for presbyopes. *Science advances*, 5(6):eaav6187, 2019.

[44] J.-H. Park and B. Lee. Holographic techniques for augmented reality and virtual reality near-eye displays. *Light: Advanced Manufacturing*, 3(1):1–14, 2022.

[45] Y. Peng, S. Choi, J. Kim, and G. Wetzstein. Speckle-free holography with partially coherent light sources and camera-in-the-loop calibration. *Science advances*, 7(46), 2021.

[46] Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein. Neural holography with camera-in-the-loop training. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020.

[47] K. Rathinavel, H. Wang, A. Blate, and H. Fuchs. An extended depth-at-field volumetric near-eye augmented reality display. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 24(11):2857–2866, 2018.

[48] K. Rathinavel, G. Wetzstein, and H. Fuchs. Varifocal occlusion-capable optical see-through augmented reality display based on focus-tunable optics. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 25(11):3125–3134, 2019.

[49] L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik. Towards real-time photorealistic 3d holography with deep neural networks. *Nature*, 591(7849):234–239, 2021.

[50] C. Su, Y. Peng, Q. Zhong, H. Li, R. Wang, W. Heidrich, and X. Liu. Towards vr and ar enhancement: light field display with mid-air interaction. In *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, pp. 1–2. 2016.

[51] K. Takahashi, Y. Kobayashi, and T. Fujii. From focal stack to tensor light-field display. *IEEE Transactions on Image Processing*, 27(9):4571–4584, 2018.

[52] M. Tatzgern, R. Grasset, D. Kalkofen, and D. Schmalstieg. Transitional augmented reality navigation for live captured scenes. In *Proc. IEEE Virtual Reality (VR)*, pp. 21–26, 2014.

[53] T. Ueda, D. Iwai, T. Hiraki, and K. Sato. Illuminated focus: Vision augmentation using spatial defocusing via focal sweep eyeglasses and high-speed projector. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 26(5):2051–2061, 2020.

[54] H. Urey, K. V. Chellappan, E. Erden, and P. Surman. State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE*, 99(4):540–555, 2011.

[55] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar. Layered 3d: Tomographic image synthesis for attenuation-based light field and high dynamic range displays. *ACM Transactions on Graphics (TOG)*, 30(4), 2011.

[56] L. Xiao, A. Kaplanyan, A. Fix, M. Chapman, and D. Lanman. Deepfocus: Learned image synthesis for computational displays. *ACM Transactions on Graphics (TOG)*, 37(6), Dec. 2018.

[57] H. Yu, M. Bemana, M. Wernikowski, M. Chwesiuk, O. T. Tursun, G. Singh, K. Myszkowski, R. Mantiuk, H.-P. Seidel, and P. Didyk. A perception-driven hybrid decomposition for multi-layer accommodative displays. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 25(5):1940–1950, 2019.

[58] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.