Gaze-Contingent Layered Optical See-Through Displays with a Confidence-Driven View Volume

Christoph Ebner, Alexander Plopski, Dieter Schmalstieg and Denis Kalkofen

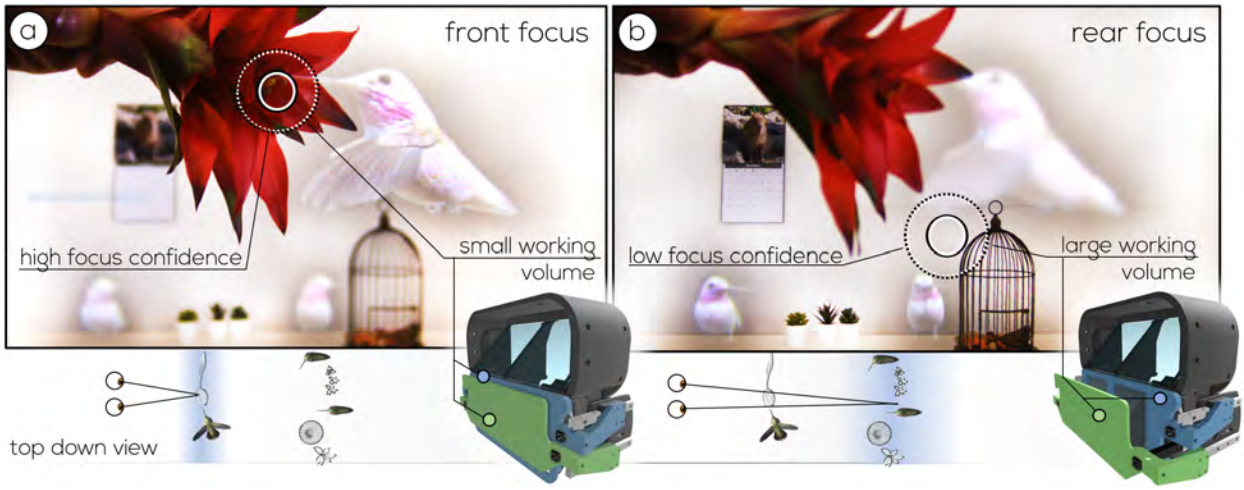

Fig. 1: We introduce gaze-contingent layered optical see-through displays to adjust the size and position of the view volume. (a) The user is focusing on the real flower in the front, which supports estimating the focus distance with high confidence and thus allows setting up a small view volume. (b) The user focuses on the back. However, the focus distance estimation is ambiguous, yielding several possible candidates for the focus distance. Thus, the selected focus distance has lower confidence. To increase the likelihood of enclosing the focus point of the user, our system adjusts by enlarging the view volume. Both pictures show photographs captured through the lens of our prototype.

Abstract— The vergence-accommodation conflict (VAC) presents a major perceptual challenge for head-mounted displays with a fixed image plane. Varifocal and layered display designs can mitigate the VAC. However, the image quality of varifocal displays is affected by imprecise eye tracking, whereas layered displays suffer from reduced image contrast as the distance between layers increases. Combined designs support a larger workspace and tolerate some eye-tracking error. However, any layered design with a fixed layer spacing restricts the amount of error compensation and limits the in-focus contrast. We extend previous hybrid designs by introducing confidence-driven volume control, which adjusts the size of the view volume at runtime. We use the eye tracker's confidence to control the spacing of display layers and optimize the trade-off between the display's view volume and the amount of eye tracking error the display can compensate. In the case of high-quality focus point estimation, our approach provides high in-focus contrast, whereas low-quality eye tracking increases the view volume to tolerate the error. We describe our design, present its implementation as an optical-see head-mounted display using a multiplicative layer combination, and present an evaluation comparing our design with previous approaches.

Index Terms—Gaze-Contingent Layered Display, Optical See-Through Mixed Reality, Vergence-Accommodation Conflict

1 Introduction

Augmented reality (AR) enables 3D computer-generated objects to be presented within the real-world environment of its user [36]. Among the existing types of AR displays, a head-mounted display (HMD) is arguably the most versatile. However, the vergence-accommodation conflict (VAC) represents a major perceptual challenge. The discrepancy between the fixed image plane of the display and the user's focus distance can cause conflicting visual stimuli [15], which increases eye fatigue and reduces performance [2, 3, 37].

To mitigate VAC, various display designs aim to increase the image contrast of objects in focus [17, 19]. For example, varifocal displays increase the contrast by aligning the distance of the virtual image plane to match that of the user's focusing distance at runtime. However, due to inherent eye tracking errors, focus estimates may be incorrect, resulting in a mismatch between user focus and image plane placement [9], causing a loss of in-focus contrast that is proportional to the error.

Layered light field displays do not need eye tracking because they present the light field between its layers at once [26] by using combinations of pixels from adjacent layers. However, the number of possible pixel combinations is restricted by the resolution of the display panels, so layered displays can provide only a compressed representation of the light field. Compression reduces in-focus contrast, a problem that increases with the size of the view volume between two layers. To maintain a small view volume over a wide viewing range, Ebner et al. [12] introduced gaze-contingent layered displays, which combine the characteristics of layered and varifocal displays. The approach

- *Christoph Ebner and Alexander Plopski are with Graz University of Technology. E-mail: {christoph.ebner | alexander.plopski}@tugraz.at*
- *Dieter Schmalstieg is with University of Stuttgart and Graz University of Technology. E-mail: dieter.schmalstieg@visus.uni-stuttgart.de*
- *Denis Kalkofen is with Flinders University and Graz University of Technology. E-mail: denis.kalkofen@flinders.edu.au*

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

makes use of eye tracking to shift the display layers until the view volume encloses the user's current focus point. Shifting the view volume of layered displays based on the user's current focus distance can mitigate the impact of erroneous focus estimation while supporting large workspaces.

However, since the contrast changes inversely to the size of the view volume, a static volume size must necessarily represent a trade-off between the error tolerance and the amount of contrast of the display. Since the quality of focus estimation is commonly affected by the complexity of the scene and the uncertainty of the eye tracker, the predefined view volume can easily become too large or too small for the quality of the focus estimator at runtime. For example, if the focus estimation performs better than assumed, the view volume is set too large, and, consequently, supported contrast is lower than it could be. If the focus estimation performs worse than assumed, the view volume might not be large enough to compensate for the offset between the estimated and actual focus distances.

To support the highest possible image contrast at all times, we adapt the size of the view volume to match the current quality of the focus estimation at runtime. In situations where the system has high confidence in the estimated focus depth, e.g., when the user looks at a flat surface, the layers of the display can be moved close together, resulting in a minimal view volume that provides the highest possible contrast in focus regions. When the system has low confidence in the estimated focus depth, for example, due to inaccurate eye tracking, the display layers need to be moved further apart to span a larger view volume around the estimated focus point (see Figure 1 for an example).

In summary, our work makes the following key contributions:

- We introduce a novel display architecture that is capable of dynamically adjusting the position and extend of its view volume based on an estimate of its user's current focus distance and the confidence of the estimation.
- We built a prototype device using a multiplicative two-layer design, which we integrated into an OST-HMD. The prototype was built from standard components, demonstrating the integration of all system components into a portable form factor. To support the reproducibility of our prototype, we also developed an approach for automatically calibrating a layered display with dynamic image planes.
- We present a novel approach for the fast computation of patterns using a panel-aligned focal stack, and we introduce a novel approach to compensating the offset between the assumed location of an image pattern and its actual location, which is especially useful for moving display panels.
- We analyze our design and show that it outperforms previous display architectures by achieving higher in-focus contrast and overall image quality, and we demonstrate the image quality of the prototype device with photographs captured through its lens.

2 Related Work

A traditional HMD uses a single image plane placed at a predefined distance. The mismatch between the distance of the user's focus point and the distance of the image plane is known to cause the VAC [15]. Several display designs have been proposed to mitigate VAC by supporting natural in-focus contrast [19].

2.1 Varifocal displays

An extension of the traditional HMD design allows adjustment of the depth in which virtual content is presented. This adjustment can either be accomplished by physically shifting the position of the display relative to the focusing lens [18] or by altering the power of the employed focusing lens, e.g., with Alvarez lenses [43], electrically [1, 23, 30, 35] or pressure-refocusable lenses [10]. In some cases, the focal depth of only a portion of the display is adjusted to ensure correct depth cues in combination with foveated rendering [18].

2.2 Multifocal displays

While varifocal displays continuously align a single image plane with the user's focus distance, the multifocal design uses multiple image planes placed at varying distances from the user [15]. When virtual content is rendered at a depth that matches one of the image planes, it can be presented on the corresponding display. If the depth of the virtual content lies between two image planes, natural focusing cues can be recovered by rendering it proportionally on both planes. This weighted approach results in more natural focusing cues than discrete rendering on the closest image plane. However, these approaches lead to artifacts at depth discontinuities and require optimized routines for decomposing a focal stack onto the layers [28, 29]. MacKenzie et al. showed that, for multifocal displays, focal planes must be placed within 0.6-0.9 dpt of each other to mitigate VAC [24]. Several variations of multifocal displays have been explored, including stacking of beam splitters [4], deformable mirrors [10], and combinations of electrically tunable lenses with displays that provide a high refresh rate [6, 34].

2.3 Light field displays

Light field displays aim to reproduce the light distribution of a virtual scene. A common approach to recovering this representation is the use of microlenses [20]. Although this approach allows for a thin display design, it limits the display resolution. Another common approach to generating a light field is to encode the arrangement of multiple display layers along the line of sight [16, 25, 41]. Here, light rays are seen as a multiplication of pixels from all displays. Besides the loss of brightness due to stacking of LCD panels, a common issue with this approach is the contrast loss when the user focuses in-between the display planes.

2.4 Dynamic layer placement

Distributing the planes of a layered display uniformly results in a loss of contrast. Wu et al. [44] presented an approach that determines the optimal placement of a finite number of focal planes based on the content in the scene. However, they did not consider user focus. Ebner et al. [12] introduce eye tracking to layered displays to adjust the placement of the view volume at runtime. However, they do not adjust the size of the view volume. Later, Ebner et al. [11] present an approach to dynamically adjusting the geometry and the extent of the view volume of a layered display, which combines a direct view display, i.e., an off-the-shelf computer monitor, with a single-layer HMD. While their approach demonstrates the feasibility of a layered display with dynamic view volume, it is restricted to desk-sized workspaces.

2.5 Holographic displays

Similar to light field displays, holographic displays recover the 3D geometry of a target object. A spatial light modulator (SLM) is used to control the phase amplitude of a collimated laser beam so that the propagated light forms the desired image at the viewpoint [33]. Utilizing an SLM and a collimated laser in holographic displays results in a complicated optical setup with a limited field of view. Holographic displays require arrangements of optical elements that often cannot be replicated in a wearable form factor [5]. In addition, achieving high-quality holographic images is computationally expensive, and noise and color discontinuities can affect the image. In recent years, the use of neural networks to compute source modulation has reduced computational demands [32]. However, rendering a single image can still take several seconds.

3 Overview

We present an optical see-through head-mounted display (OST-HMD) that provides a confidence-driven view volume. The view volume determines the depth range in which the focus cues can be effectively provided to the users. A larger view volume is desirable, as it enables the presentation of content across a wider depth of field without causing visual discomfort. However, as the view volume expands, the perceived contrast for objects located between the two layers decreases. Thus, our objective is to strike a balance between the view volume and the quality of focus cues in terms of contrast. Therefore, our layered display consists of two stacked displays, each of which can be controlled by a varifocal mechanism. This allows for independently moving each layer and to dynamically adjust the view volume (the volume bound by the two display layers). Refer to Figure 2 for an overview of this process.

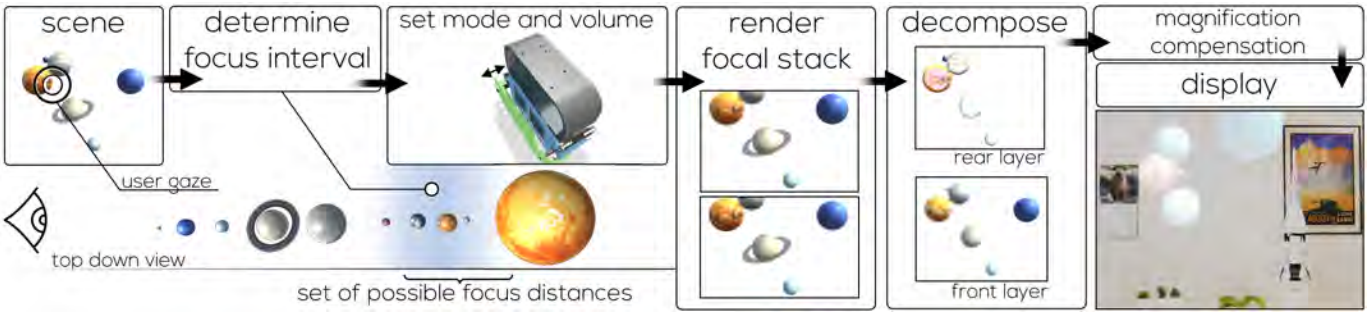

Fig. 2: We render an all-in-focus image of the scene and concurrently determine the set of possible focus distances for the user. The upper and lower bounds of this set determine the current view volume and the range of the focal stack. We select a display mode (varifocal or layered) according to the extent of the view volume and shift one or both of the physical layers to setup the view volume. For the layered mode, we decompose the focal stack to get attenuation patterns to show on the LCD panels. Finally, we post-process the images shown on the LCD panels and display the result.

To achieve near-optimal contrast, we dynamically adjust the view volume based on the estimated focus distance of the user. Specifically, we determine the depth interval in which the user focus distance is likely to be located (Section 3.1). The two display layers that limit the view volume are then set to the minimum and maximum of this range (Section 3.2). In other words, the extent of the view volume reflects the confidence in the focus distance estimation. If, for example, the range is large, i.e., if there is low confidence in the focus distance estimation, the view volume also has a large extent. Conversely, with increasing certainty of the estimated focus distance, we can reduce the view volume. In highly certain cases, we can disable a single layer and use the display in a varifocal mode, e.g., when the user is looking at a wall that has a label rendered on it.

Existing layered display implementations make use of an additive or multiplicative image formation model. Additive blending of pixels in different layers is commonly achieved with beam splitters [15], while multiplicative pixel formation is often implemented by stacking LCD panels [16, 41, 42]. Since additive blending lacks support for pixel-perfect occlusions of arbitrary scene geometry [7], and the spatial separation of display panels commonly results in bulky displays, we chose to build our prototype based on multiplicative image formation. In particular, we use a stack of two LCD panels, which we position so that the image planes enclose the focus point of the user. However, since high-resolution multiplicative layered displays suffer from diffraction of light passing through the front panel, we combine a low-resolution front with a high-resolution back panel. The use of a front display panel with lower resolution allows for setting up a reasonable sized volume without degrading the back panel resolution.

We drive the displays by rendering a dense focal stack within the estimated range, which we use as ground truth for the output at given focus distances (Section 3.3). To approximate a light field with our layered display, we decompose the focal stack into two attenuation patterns (Section 3.4).

As a by-product of its design, our display can emulate either a layered display, similar to that of Huang et al. [16], or a varifocal display (by making one of the layers transparent). For the special case where the system confidence is high, but not high enough to switch to a varifocal design, we introduce a novel approach to the decomposition of a panel-aligned focal stack (Section 3.5).

During computation of the attenuation patterns, we need to make an assumption of where the virtual images of the display layers are currently located. However, the layers might shift during the decomposition process, e.g., as the user refocuses. This leads to misaligned pixels of the layers which prevents emitting the light field correctly. To compensate for this effect, we introduce a transformation that maps the patterns onto their corresponding display planes in their current configuration while preserving the intended pixel combinations (Section 3.6). Our transformation is also capable of compensating for offsets that result from moving display planes. For example, when repositioning a display panel is slower than the refresh rate of the display, the system must compensate for the offset between the target and the image location at refresh time. Our approach involves a series of lookup tables filled during display calibration. For self-contained operation, we use an automatic calibration approach that captures images through the lens of the display with a camera, while the display adjusts the position and size of the view volume (Section 3.7).

3.1 Determining the focus interval

Several strategies exist to infer the current focus distance of users, including autorefractors [27] and eye tracking solutions [14]. In our system, we rely on video-based eye tracking, one of the most common methods of gaze prediction.

Within this framework, previous work has demonstrated two distinct approaches for focus distance measurement: The first method utilizes a depth map and infers the focus distance by intersecting the estimated gaze point with the depth map, leveraging the assumption that users focus on the objects they are gazing at [8]. The second method capitalizes on the natural coupling of vergence and accommodation, indirectly estimating the focus distance by measuring the vergence angle of the user's eyes [22]. Both methods can be combined as a sensor fusion approach [31, 40]. However, current video-based eye tracking systems exhibit an error of approximately 1° [9]. Consequently, both methods are susceptible to errors and have uncertainties in the predicted focus distance. To our knowledge, we are the first to incorporate the predicted focus distance *with* the associated uncertainty of the measurement.

We determine the view volume using the following steps (refer to Figure 3): We calculate the eye tracking uncertainty in diopters ε_{dpt}, given the eye tracker accuracy in degrees ε_{deg}. The uncertainty can be determined using the calibration points of the eye tracker or, alternatively, as described by Dunn [9, eq. 8]. For example, an average IPD of 63 mm and an accuracy of 1°, 2°, and 3°yields an eye tracker uncertainty of about 0.3 dpt, 0.6dpt, and 0.9 dpt, respectively. This uncertainty is used in conjunction with the vergence distance u_v (measured by the eye tracker at runtime) to define a rough range $\mathcal{V} = u_v \pm \varepsilon_{\text{dpt}}$ in which the true focus distance of the user might be. Within this depth range, each depth value d is assigned a weight using a Gaussian function centered at u_v with the parameter ε_{dpt}. Finally, we normalize the weights; depth values outside of the allowed range are assigned a weight of zero.

Additionally, we intersect the depth map of the renderings with the estimated gaze point u_g of the user that is provided by the eye tracker. To address the accuracy of the eye tracker, we define a cone with the predicted gaze direction as the axis and an opening angle proportional to the eye tracker uncertainty. Inside the cone, we collect a second set of possible candidates for the user's focus distance $\mathcal{W}$. Similarly to the first step, we propose a 2D Gaussian weighting of the depth candidates. If two or more candidates with the same depth values are contained in $\mathcal{W}$, we add their weights before normalization.

Finally, both sets are combined to form a set $\mathcal{U}$ with multiplied weights per depth value. In case two sets $\mathcal{V}$ and $\mathcal{W}$ do not have overlapping depth values, we choose $\mathcal{U}$ so that it corresponds to the set with a smaller depth variance, suggesting a higher confidence of the predicted focus distance.

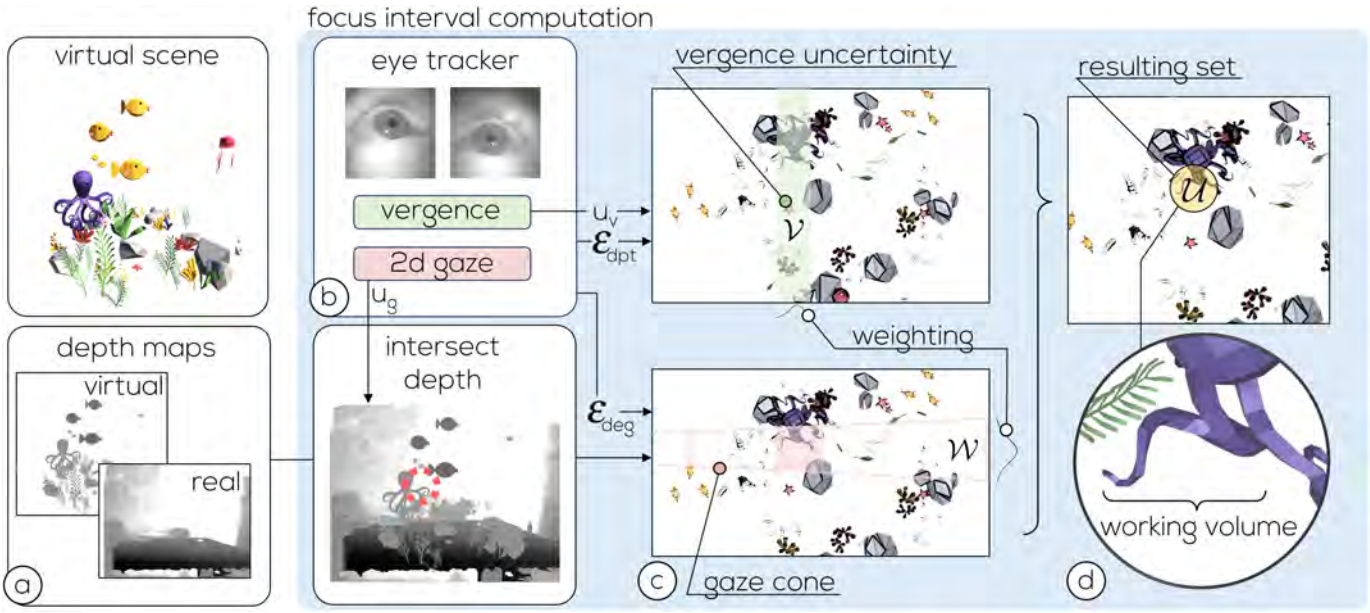

Fig. 3: Calculating the interval that contains possible focus distances of the user. (a) We combine the depth maps of the rendering with the depth map of the real world to create a joint depth map, which we later use for gaze intersection. (b) We use the vergence and 2D gaze estimation from the eye tracker in conjunction with the eye tracking uncertainty as input for focus distance estimation. (c) The vergence distance and the uncertainty lead to a range of possible depth candidates for focus distance estimation (top), while the intersection of the gaze point with the depth map and the uncertainty in degrees lead to a cone containing the second set of possible candidates (bottom). (d) We weigh both sets and combine them into a resulting set which consists of possible user focus distances.

3.2 Setting up the view volume

Depending on the extent of the resulting set $\mathcal{U}$, we set the view volume of the display. If the extent of $\mathcal{U}$ is small, our system also offers the option of making one layer transparent to effectively use the HMD as a varifocal display. In this case, the view volume – per definition – is zero. In contrast, when the extent of $\mathcal{U}$ is substantial, a wide range of potential focus distances must be supported. Consequently, both layers of the display are enabled to emit a light field to support the broader range of possible focus distances. In this case, the view volume is delineated by the depth range that encompasses the 95^{th} percentile of the $\mathcal{U}$ distribution. To establish suitable thresholds for selecting the mode and view volume, we performed a contrast analysis to select the mode and view volume that yield the optimal contrast in each situation (refer to Section 5).

3.3 Focal stack rendering

Input to our display are renderings with natural depth of field. In varifocal mode, this simply entails rendering the scene with depth of field corresponding to the inferred focus distance of the user. In layered mode, we render a dense focal stack within the view volume. The generated focal stack consists of N ground-truth images corresponding to the potential user focus distances and will be utilized for the subsequent decomposition process that yields the display attenuation patterns.

We render depth-of-field images using the approach of Ebner et al. [12], which is capable of generating a dense focal stack in real time. To ensure that the rendered focal stack images closely resemble the user's perception, we approximate the human eye using a camera model comprising an ideal thin lens and a planar image sensor with a resolution of $X \times Y$ pixels. The distance from the lens to the sensor is set to $d_s = 17$ mm. We adjust the aperture size of the camera used in the focal stack rendering to match the average pupil size, $A = 4$ mm.

Note that the aperture size determines the eye box of the light field emitted by the display. Huang et al. [16] chose an eye box that was larger than the pupil to account for eye movements. However, this approach decreases the contrast of objects not located on the image plane, as their circle of confusion on the image plane increases. Additionally, a larger aperture incurs higher rendering and decomposition costs. With the help of the eye tracker, any eye movement can be translated into a simple camera movement in the renderer, eliminating the need to extend the eye box beyond the pupil size.

3.4 Obtaining attenuation layer patterns

For computing the display images, we must decompose the focal stack into the patterns to be shown on the panels. Previous work has demonstrated the use of focal stacks for light field reconstruction, but these methods were limited to focal stacks that contain images with focus distances corresponding to the virtual image planes of the display [38]. To maintain a high refresh rate, we decouple the decomposition from the display of the resulting image patterns.

Input to the decomposition scheme is a dense focal stack. Each image in the focal stack, denoted as $\hat{\boldsymbol{r}}_n$, $n \in \mathbb{N}, 0 < n \leq N$, is weighted using the depth weights in $\mathcal{U}$ to incorporate the likelihood that the user will focus at a particular distance during the decomposition process. If an image in the focal stack does not directly correspond to a depth value, such as in cases where no depth value is available for the given focus distance within the set of possible candidates, we perform a linear interpolation on the weights of the two adjacent depth values and utilize the resulting weight w_n for that image. To balance the overall brightness per pixel, the sum of these weights is normalized to one.

To predict the perceived retinal image when looking through the display, we sample the light field emitted by the display across the pupil. The number of views necessary to generate an image of the focal stack without aliasing depends on the focus distance d_f, the distance of the displays d_i, the display resolution r_i and the current extent of the virtual image plane e_i. The number of views M_n is calculated as

$$M_n = 4 \left\lceil \max_{i \in \{1,2\}} A \frac{|d_i - d_f| \cdot r_i}{d_f \cdot e_i} \right\rceil^2. \tag{1}$$

Defining the pixels of the two display panels with resolutions $U \times V$ and $S \times T$ as vectors $\boldsymbol{p}_1 \in \mathbb{R}^{UV}$ and $\boldsymbol{p}_2 \in \mathbb{R}^{ST}$, respectively, the perceived retinal image vector $\boldsymbol{r}_n \in \mathbb{R}^{XY}$ when looking through the display amounts to

$$\boldsymbol{r}_n = \boldsymbol{R}_n \cdot \left(\boldsymbol{A}_{n,1} \cdot \boldsymbol{p}_1 \odot \boldsymbol{A}_{n,2} \cdot \boldsymbol{p}_2 \right), \tag{2}$$

where the symbol $\odot$ refers to the Hadamard product, the matrices $\boldsymbol{A}_{n,1} \in \mathbb{R}^{M_n XY \times UV}$ and $\boldsymbol{A}_{n,2} \in \mathbb{R}^{M_n XY \times ST}$ map the individual pixels to the image sensor for the individual views M_n, and the matrix $\boldsymbol{R}_n \in \mathbb{R}^{XY \times M_n XY}$ generates an image with the corresponding focus distance from the emitted light field. To generate the display patterns, we formulate the following objective function:

$$\underset{\boldsymbol{p}_i,\, i\in\{1,2\}}{\arg\min} \; || \sum_{n=1}^{N} w_n \big(\hat{\boldsymbol{r}}_n - \boldsymbol{r}_n\big) ||, \quad \text{s.t. } 0 \leq \boldsymbol{p}_i \leq 1. \tag{3}$$

To find appropriate display patterns using Equation 3, we employ non-negative tensor factorization [38,42]. We initialize both patterns to be fully transparent and perform the following update steps in alternating manner, until convergence:

$$\begin{aligned} \boldsymbol{p}_1^{(m+1)} &= \boldsymbol{p}_1^{(m)} \odot \sum_{n=1}^{N} w_n \frac{\boldsymbol{F}_{n,1}^T(\hat{\boldsymbol{r}}_n \odot (\boldsymbol{F}_{n,2} \cdot \boldsymbol{p}_2^{(m)}))}{\boldsymbol{F}_{n,1}^T(\boldsymbol{r}_n \odot (\boldsymbol{F}_{n,2} \cdot \boldsymbol{p}_2^{(m)}))}, \\ \boldsymbol{p}_2^{(m+1)} &= \boldsymbol{p}_2^{(m)} \odot \sum_{n=1}^{N} w_n \frac{\boldsymbol{F}_{n,2}^T(\hat{\boldsymbol{r}}_n \odot (\boldsymbol{F}_{n,1} \cdot \boldsymbol{p}_1^{(m+1)}))}{\boldsymbol{F}_{n,2}^T(\boldsymbol{r}_n \odot (\boldsymbol{F}_{n,1} \cdot \boldsymbol{p}_1^{(m+1)}))}, \end{aligned} \tag{4}$$

where $\boldsymbol{F}_{n,i} = \boldsymbol{R}_n \, \boldsymbol{A}_{n,i}$. After each update, the patterns are clamped to the range $[0;1]$.

3.5 Decomposition with panel aligned focal stack images

A special case exists in which the number of focus distances in the set $\mathcal{U}$ is large enough to require two display layers, but small enough to only incorporate two images in the focal stack. The focus distances of these images correspond to the distances of the display's image planes and, thus, to the boundaries of the view volume. In this case, the light field decomposition problem is not overdetermined anymore, as the number of images in the focal stack corresponds to the number of display layers. Further, the point spread function when focusing to the distances corresponding to the layers reduces to a point. This special case enables one to update the layers using a special scheme (refer to the supplemental material for a rigorous derivation of this equation):

$$\tilde{\boldsymbol{p}}_i^{(m+1)} = \tilde{\boldsymbol{p}}_i^{(m)} + \frac{1}{2}\left(\tilde{\boldsymbol{r}}_i - (\tilde{\boldsymbol{p}}_i^{(m)} + \tilde{\boldsymbol{p}}_j^{(m)} * \boldsymbol{c}_{i,j})\right), \tag{5}$$

where $\tilde{\boldsymbol{p}}_i$ and $\tilde{\boldsymbol{r}}_i$ refer to the log-transformed pattern and focal stack, respectively. $\boldsymbol{c}_{i,j}$ denotes the circle of confusion on layer i, when focusing on layer j, where $j = 3 - i$. Assuming that the panels are initialized with zeros, the first iteration maps each focal slice to its corresponding panel with half intensity, i.e., $\tilde{\boldsymbol{p}}_i = \tilde{\boldsymbol{r}}_i/2$. Equation 5 lets us derive a closed-form solution for iteration m, which does not depend anymore on the previous iteration. Thus, the patterns for each panel can be directly obtained from the focal stack in a non-iterative manner as

$$\tilde{\boldsymbol{p}}_i^{(m)} = \sum_{k=1}^{\lceil m/2 \rceil} \eta_k \cdot \tilde{\boldsymbol{r}}_i \overset{k-1}{*} \boldsymbol{c}_{j,i} \overset{k-1}{*} \boldsymbol{c}_{i,j} - \sum_{k=1}^{\lceil m/2 \rceil} \zeta_k \cdot \tilde{\boldsymbol{r}}_j \overset{k-1}{*} \boldsymbol{c}_{j,i} \overset{k}{*} \boldsymbol{c}_{i,j}, \tag{6}$$

where $g_0 *^k g_1$ denotes k-fold convolution of a function g_0 with the function g_1, i.e., $g_0 *^k g_1 = g_0 \underbrace{* g_1 * g_1 \cdots * g_1}_{k}$, and m is referred to as decomposition order. Finally, the weighting functions η_k and ζ_k for the focal images $\tilde{\boldsymbol{r}}_i$ and $\tilde{\boldsymbol{r}}_j$ are defined as:

$$\eta_k = \frac{1}{2^m}\left(2^m - \sum_{\mu=0}^{2(k-1)} \binom{m}{\mu}\right), \quad \zeta_k = \frac{1}{2^m}\left(2^m - \sum_{\mu=0}^{2k-1} \binom{m}{\mu}\right), \tag{7}$$

where $\binom{m}{\mu}$ refers to the binomial coefficient. Furthermore,

$$\sum_{k=1}^{\lceil m/2 \rceil} (\eta_k - \zeta_k) = 0.5. \tag{8}$$

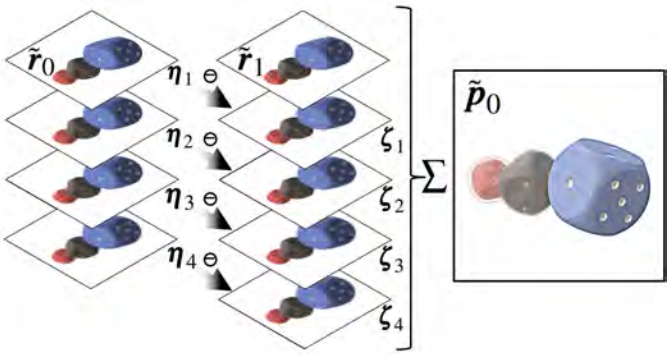

Fig. 4: Visual representation of Equation 6, depicting the initialization approach for the display transition. The computation of the pattern $\tilde{\boldsymbol{p}}_0$ for a decomposition of order $m = 8$. During computation, differently blurred versions of the images in the focal stack are subtracted from each other, and the result is added to produce the pattern.

Note that after using Equation 5 to compute the log-transformed patterns, we transform the resulting patterns back using the exp function: $\boldsymbol{p}_i = e^{\tilde{\boldsymbol{p}}_i}$. While Equations 5 and 6 are equivalent in a mathematical sense, the latter gives an intuition of what happens during the decomposition process. For an increasing order of decomposition m, the focal slices are convolved with an increasing number of blur kernels $c_{k,j}$ and $c_{k,i}$ that are subtracted from each other. Note that since Equation 6 is no longer iterative, the pattern values are not clamped to the interval $[0;1]$ after each iteration. However, to prevent divergence, Equation 6 can be used as an initialization scheme to push the decomposition process towards a converged solution that is followed by a few iterative steps. We analyze the impact of this initialization algorithm on runtime and quality in Section 5.

Equation 6 lends itself to performance optimizations to reduce runtime. First, since the scheme is no longer iterative, each step of the sum in Equation 6 can be computed independently by a separate processor. Subsequently, the results can be shared and added to yield the layer pattern. The increase in computational cost for higher decomposition orders due to the accumulation of convolution kernels can be tackled by performing the optimization in the Fourier domain. Alternatively, the convolutions can be computed using a summed area table (SAT). Once obtained, the SAT enables computing the circular convolution through multiple box blurs. The computation of the SAT itself can be performed in $\mathcal{O}(\log n)$ time with parallelization [13].

Intuitively (and assuming the point spread function is Gaussian), the panel decomposition scheme of Equation 6 can be thought of as building a Gaussian pyramid for each image in the focal stack and subtracting adjacent layers from the two pyramids (Figure 4). Thus, each panel is comprised of a weighted sum of the difference of Gaussians per slice, as well as the differences due to the varying focal distances across focal images. Using a slightly different interpretation, the decomposition implicitly sums up the layers of a Laplacian pyramid, in which every other layer consists of a differently blurred version of the same image of the focal stack. With higher decomposition orders, blurrier versions of the focal images are subtracted from each other. However, these higher-order terms only play a minor role compared to the lower-order terms, as weights decrease for upper pyramid layers. As shown in Figure 4, the resulting patterns exhibit pronounced ring structures, which have been attributed in the past as crucial to driving accommodation in multifocal displays [28].

3.6 Pattern alignment

Once the view volume of the display has been defined, we shift the display panels to the locations corresponding to the calculated view volume. However, it is important to note that any modification of the distance of the virtual image plane will result in a change in magnification. Such a change to the field of view is a common occurrence in displays employing a varifocal mechanism. In our system, this poses an

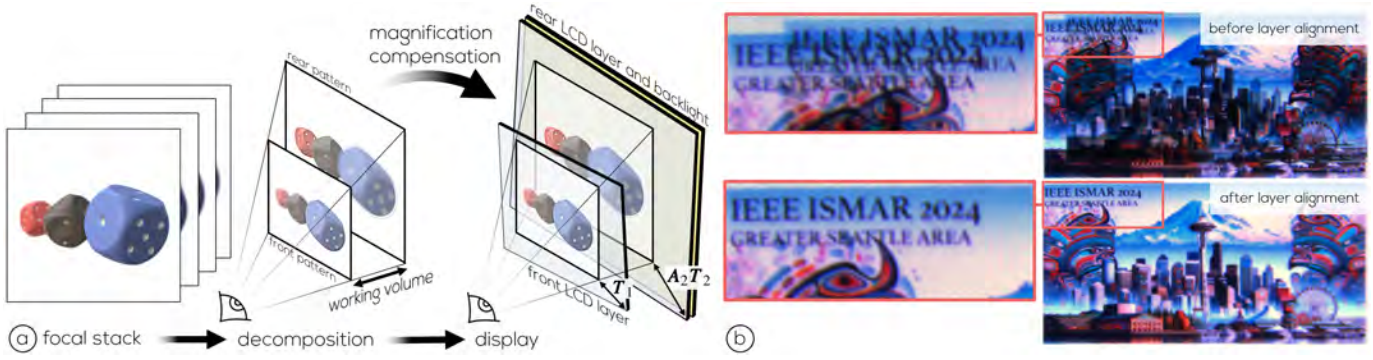

Fig. 5: Alignment. (a) We use a dense focal stack containing images that are differently focused within the working volume as input to the decomposition. The decomposition does not make any assumptions about the current extent of the virtual image of the LCD layers. This allows us to decompose the scene concurrently with moving the LCD layers without needing to know beforehand in which position they end up when the decomposition finishes, thereby reducing latency. For display, the patterns are subsequently transformed via the pre-calibrated transforms $\boldsymbol{T}_i$ and $\boldsymbol{A}_i$ to be aligned again. As indicated in the figure, in addition to compensating for different magnifications, these transforms also work for differently rotated and tilted displays. (b) Impact of computing $\boldsymbol{A}_2$ on the alignment of the display layers. (top) Through-the-lens view showing unaligned layers and (bottom) layers aligned using our magnification compensation with the parameters we retrieve from the automatic calibration routine.

additional challenge, as any misalignment of the two layers is unacceptable for a light field display. Therefore, it is necessary to compensate for the change in magnification within the system (see Figure 5 for an illustration). It is crucial that this compensation is executed in an imperceptible manner to prevent users from experiencing flickering.

To compensate for the change in magnification, we perform an automatic calibration routine in which we calculate lookup tables to adjust the displayed content at runtime. To this end, we define a reference configuration that is determined by the lower bound of the field of view. For a single display panel, the smallest extent of the virtual display image is generated with either the closest distance to the lens (in case of a mechanical shifting mechanism) or the highest focal power of the lens (in case of a tunable lens). Subsequently, we refer to the distance closest to the lens of the layer i as its reference distance $\delta_{i,ref}$, $i \in \{1,2\}$, and the reference distance of the front panel $\delta_{1,ref}$ as the reference configuration of the whole system.

To address changes in field of view resulting from different distances between the panel and the lens, we utilize a linear image transform $\boldsymbol{T}$ that adjusts the displayed image to a smaller field of view, given as

$$\boldsymbol{p}_i = \boldsymbol{T}_i(\delta_i) \cdot \tilde{\boldsymbol{p}}_i, \tag{9}$$

where $\boldsymbol{T}_i(\delta_i)$ is a 2×3 matrix that depends on the desired distance of the virtual image plane, $\tilde{\boldsymbol{p}}_i$ represents the pixels displayed on panel i before the shift, and $\boldsymbol{p}_i$ represents the pixels displayed after the virtual image plane has been adjusted. Note that, in addition to the change in magnification, $\boldsymbol{T}_i(\delta_i)$ accounts for deviations to the reference pose. This equation models a transform of the display to account for different shifts of the virtual image plane. In addition to $\boldsymbol{T}_i(\delta_i)$, which transforms a single layer at position δ_i into its reference position, we use another transform $\boldsymbol{A}_i$ that aligns the reference configuration of the rear layer to the reference configuration of the front layer (i.e., $\boldsymbol{A}_1 = \boldsymbol{I}$), making sure that the pixels of the two layers are properly aligned. Figure 5a depicts the process of aligning the pixels for display.

Lastly, we introduce a color-dependent aberration function $\boldsymbol{D}_i(\tilde{\boldsymbol{c}}_i)$ that models HMD lens aberrations, such as distortion and axial chromatic aberration. In our prototype, the focal power of the lens does not change; thus, we use a single $\boldsymbol{D}_i$ for each color channel. If the varifocal mechanism is established through a focus-tunable lens, it is straightforward to make $\boldsymbol{D}_i$ dependent on the current focal power.

3.7 Calibration

To calibrate $\boldsymbol{T}_i(\delta_i)$, $\boldsymbol{A}_2$ and $\boldsymbol{D}_i$, we extend the approach of Lee and Hua [21] to layered displays. We automate the calibration by positioning a camera in front of the HMD, approximately at the location where the user's eye can be assumed, and set the virtual image plane to the reference configuration. A checkerboard is displayed in the HMD and observed by the camera. We use the corners of the checkerboard pattern to establish a relationship between the pattern displayed on the layer ($\tilde{\boldsymbol{c}}$) and the corresponding pattern in the camera image $\boldsymbol{c}_i$,

$$\boldsymbol{c}_i = \boldsymbol{K} \cdot \boldsymbol{P}_i \cdot \boldsymbol{A}_i \cdot \boldsymbol{T}_i(\delta_i) \cdot \boldsymbol{D}_i(\tilde{\boldsymbol{c}}), \tag{10}$$

where $\boldsymbol{K}$ represents the intrinsics of the camera, and $\boldsymbol{P}_i$ represents the pose of the virtual image plane in camera space. Calibration starts by computing $\boldsymbol{P}$ and $\boldsymbol{D}$ for the front layer in the reference configuration, where $\boldsymbol{T}_1(\delta_{ref}) \equiv \boldsymbol{I}$, by solving the optimization problem

$$\underset{\boldsymbol{P}_1, \boldsymbol{D}_1}{\arg\min} ||\boldsymbol{c}_1 - \boldsymbol{K} \cdot \boldsymbol{P}_1 \cdot \boldsymbol{D}_1(\tilde{\boldsymbol{c}})||, \tag{11}$$

using a predetermined camera matrix $\boldsymbol{K}$. After obtaining $\boldsymbol{P}$, we sample the depth range by shifting the display to various depth values and compute $\boldsymbol{T}$ in a similar manner as in Equation 11, with fixed $\boldsymbol{P}$ and $\boldsymbol{D}$. This process allows us to calculate $\boldsymbol{T}$ for different distances, covering the entire depth range of the layer. We then perform a polynomial fitting procedure on the scaling and translation components of the $\boldsymbol{T}_1$ matrices to obtain a continuous model that accounts for shifts in the virtual image plane, providing a smooth representation of the transformation throughout the depth range.

Finally, we perform the same routine for the rear panel. The only matrix left to determine is $\boldsymbol{A}_2$, which we compute in the reference configuration of the rear panels ($\boldsymbol{T}_2 \equiv \boldsymbol{I}$) using the camera points $\boldsymbol{c}_1$ observed on the front panel. Finally, $\boldsymbol{A}_2$ is obtained by solving

$$\underset{\boldsymbol{A}_2}{\arg\min} ||\boldsymbol{c}_1 - \boldsymbol{K} \cdot \boldsymbol{P} \cdot \boldsymbol{A}_2 \cdot \boldsymbol{D}_2(\tilde{\boldsymbol{c}})||. \tag{12}$$

The impact of aligning the layers by $\boldsymbol{A}_2$ is shown in Figure 5b.

4 Prototype

Our prototype incorporates a mechanical varifocal mechanism comprising four Actuonix P8-ST microstepper actuators, each pair responsible for controlling the movement of a single display panel. These actuators allow translating the virtual image of each LCD to arbitrary positions within 3 dpt to infinity in about one second, which roughly corresponds to the accommodation time of the human eye [18], including latency. To ensure stability during operation, we have employed two guide rails for each image plane. To accurately predict the position of each display panel, the prototype incorporates endstops. At the beginning of each session, the actuators initiate a backward movement of the display panels until the endstops are triggered. Refer to Figure 6a for a depiction of the components of the display.

For the back layer panel, we used two Sharp LS029B3SX02 displays, each with a size of 2.9 inches and a resolution of 1440×1440 pixels

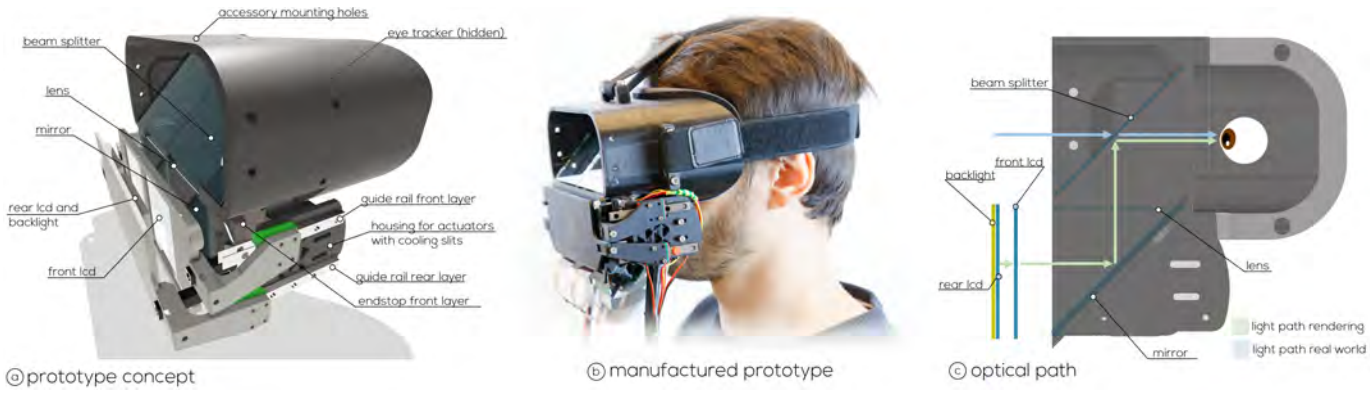

Fig. 6: Prototype. (a) A mechanical translation mechanism controls the two layers independently. For simplicity, only the right side of the display is annotated. (b) A photograph of the prototype. (c) A cross-section view of the prototype showing the optical path of the rendered scene (green color) and the real world (blue color). The light emitted by the backlight passes through the rear and front LCD panels. Subsequently, the light is reflected by a mirror and focused by a lens. A beam splitter combines reflected light rays with light entering the display from the real world.

per eye (706 ppi). These displays operate at a refresh rate of 120Hz. The front layer panel consists of two JDI LT031MDZ4000 displays, offering a resolution of 720×720 pixels per eye (329 ppi) and operating at a refresh rate of 60Hz. A deliberate decision was made in favor of a lower pixel density display for the front panel, reducing the occurrence of diffraction effects that may arise from front displays with high pixel density (see Section 5.1).

To create a virtual image using the LCD layers, we use Fresnel lenses with a focal length of 7 cm. We use a mirror between the lens and the front display to fold the optical path, reducing the overall footprint. Moreover, we employ a commercial 50/50 beam splitter to combine the light received from the real environment with the renderings. Virtual content can be displayed with a field of view of about 46°. Figure 6b presents an image of the prototype, while Figure 6c shows a cross-section of the display with light paths.

Since the front panel operates at a refresh rate of 60Hz, the system has a time interval of 16 ms for focal stack rendering, decomposition, and display. Providing new frames every 16 ms is an unrealistic goal for our current hardware setup due to the computational demand of the decomposition. However, this problem is alleviated by the fact that the decomposition is independent of the current position of the displays, as intermediate decomposition results can be displayed at every vertical synchronization (using magnification compensation), which reduces perceived flickering and latency.

5 Evaluation

We evaluate our approach by comparing simulated results to other display architectures, and by providing photographs captured through-the-lens of the prototype. The results were generated with a desktop PC (CPU: AMD Ryzen 9 3900X at 4 GHz, 64 GB RAM, GPU: NVIDIA RTX Ada 6000).

5.1 Tradeoff between view volume and resolution

Similarly to other transmissive multiplicative layered displays [16], the view volume of our system is limited by optical diffraction. This effect forces a trade-off between the view volume and the resolution of the layers. Figure 7 shows this relationship for several pixel densities of the front and rear panels. Our goal was to create a display capable of spanning a volume of at least 0.6 dpt, which is needed to account for the uncertainty in vergence measurements of current-generation eye trackers [9]. However, we did not want our display resolution to be too low, despite using commercially available LCD panels. As a consequence, our prototype supports a view volume of 0.66 dpt (highlighted in Figure 7a) if the front layer is located at 4 dpt. However, the view volume decreases as the two layers shift toward the back (Figure 7b). Since the VAC is only noticeable for close objects, the reduced view volume for distant layers is a negligible drawback.

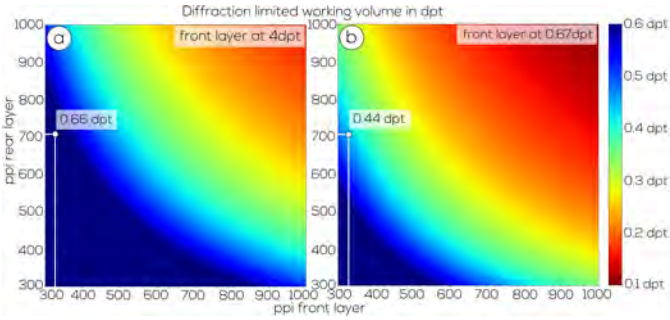

Fig. 7: Diffraction-limited view volume depending on the pixel density of the layers. Our system supports 0.66 dpt for close layers (a), and 0.44 dpt for layers located far from the user (b). In these intervals, diffraction does not degrade the perceived resolution of the rear layer.

5.2 Contrast

The display's ability to produce high contrast is important to mitigate the VAC. Therefore, we evaluate the perceived contrast by simulating display results in the event of focus distance uncertainties. We test the impact of several uncertainties that cause specific view volumes around the measured distance.

For each uncertainty setting, we render sine gratings from $1-20$ cpd at the ground truth focus distance. In addition, we simulate perceived images in the event of different offsets ε_u between the estimated focus distance and the ground truth distance. For each uncertainty σ_u (corresponding to a view volume of $2\sigma_u$ centered on the inferred focus distance), we simulate different offsets in the range of $\varepsilon_u \in [-1.5\sigma_u; 1.5\sigma_u]$.

Figure 8 shows the average perceived relative contrast depending on the uncertainty of the focus distance for our display architecture, a conventional display (single plane at 0.5 dpt), a light field display with static layers [16] (referred to as "lf stereoscope"), and a varifocal display. The results indicate that the varifocal display produces the best contrast in the event of focus distance uncertainties up to 0.08 dpt. Therefore, in this range, we chose to use the varifocal mode in our system. Furthermore, for uncertainties up to 0.125 dpt, we make use of the special decomposition with panel-aligned focal stacks.

5.3 Image quality

We further evaluate our approach in terms of image quality using three synthetic scenes of high depth complexity (see the supplemental material for images). For each scene, we simulate results at several focus distances, and we compare the simulated perceived user image for each given focus distance to the respective ground-truth image.

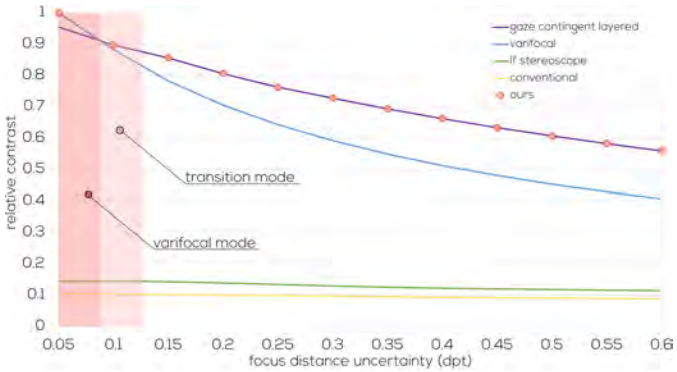

Fig. 8: Relative contrast of several display types. Our system offers the unique ability to switch from dynamic light field to varifocal mode, to maximize display contrast.

Our proposed display is tested for its best- and worst-case scenario. Therefore, we simulate two different view volumes for each focus distance offset. Each volume is centered on the inferred focus distance of the user. To simulate the best case, we set the view volume size to match the focus distance offset. To simulate the worst case, we set the size of the view volume to half of the focus distance offset, which causes the user focus distance to be outside of the volume. In addition, we compare the results achieved with our approach with simulations of related display architectures. In particular, we compare the results with a static single image plane, placed at 0.5 dpt (conventional), a display with an adjustable image plane (varifocal), and a static two-layer light field display [16] (lf stereoscope). The comparison shows results in case of a user focus distance of 0.1 dpt and offsets between the actual user focus distance and the inferred focus distance ε_u.

Table 1 contains results for three focus distances (front, center, back) and three focus distance uncertainties each, in terms of mean squared error (MSE) and the structural similarity index metric (SSIM) for the aforementioned display types. Note that, similar to contrast analysis, for a given focus distance confidence interval σ_u, we sampled focus distance errors within the range of $\varepsilon_u \in [-1.5\sigma_u; 1.5\sigma_u]$. Figure 9 contains a visual comparison in a single scene using a single focus distance. For further results with additional scenes and focus distances, please refer to the supplemental material.

Figure 9 and Table 1 demonstrate that our approach yields superior results throughout all test cases. We believe this is the result of adapting the view volume and switching to a varifocal approach in the event of high confidence focus estimation. Figure 9 shows that in its best case scenario our approach produces significantly better results for all tested offsets. However, it also shows that in the worst case scenario it only generates significantly better results for the smallest distance offsets. It is important to note that the worst-case scenario is rare, and that it almost exclusively occurs when the sets $\mathcal{V}$ and $\mathcal{W}$ disagree and the wrong set has a higher confidence. Furthermore, by discarding the set $\mathcal{V}$, one could enforce that the display covers the entire scene to avoid this scenario. Explorations on how to best combine the sets $\mathcal{V}$ and $\mathcal{W}$ to consistently achieve a high-confidence focus range estimation presents an interesting avenue of future exploration.

Figure 10 shows how image quality varies with certain focus distance offsets across given view volumes. We show peak signal-to-noise ratio (PSNR), SSIM, and learned perceptual image patch similarity (LPIPS) [45] graphs for Scene (c), using a focus distance of 1 dpt. Overall, the results show that our system manages to keep the image quality relatively constant. For example, PSNR values never drop below 30 dB in the case that the actual user focus distance is within the volume spanned by the layers. Furthermore, the image quality increases as the user focus distance approaches the volume bounds. This can be expected, because the circle of confusion on the layers decreases, which has a positive effect on image contrast.

Table 1: Image quality comparison of display architectures in three scenes. For a given confidence interval σ_u, each value represents the average of the focus distance offsets within $\pm\sigma_u$.

			MSE ↓			SSIM ↑		
			front	center	back	front	center	back
0.3 dpt uncertainty	Scene a	Ours	**17.50**	**6.56**	**264.67**	**0.989**	**0.997**	**0.985**
		Varifocal	23.83	14.59	288.69	**0.989**	0.995	0.942
		Static LF	64.27	25.99	506.33	0.949	0.964	0.897
		Conventional	323.66	113.03	489.36	0.870	0.947	0.903
	Scene b	Ours	**11.40**	**49.91**	**116.58**	**0.988**	**0.981**	**0.992**
		Varifocal	11.49	62.39	269.00	**0.988**	0.978	0.985
		Static LF	89.42	359.53	396.32	0.921	0.886	0.888
		Conventional	78.57	94.36	193.30	0.928	0.964	0.985
	Scene c	Ours	**20.25**	**118.18**	**32.32**	**0.981**	**0.964**	**0.980**
		Varifocal	22.80	140.84	171.52	0.978	0.957	0.882
		Static LF	127.81	982.28	362.75	0.891	0.759	0.792
		Conventional	177.01	235.49	262.94	0.867	0.932	0.816
0.6 dpt uncertainty	Scene a	Ours	**23.94**	**12.35**	**267.04**	**0.986**	**0.994**	**0.962**
		Varifocal	37.25	17.55	500.89	0.981	0.992	0.897
		Static LF	76.47	35.63	581.41	0.942	0.960	0.882
		Conventional	325.47	115.17	597.42	0.871	0.947	0.879
	Scene b	Ours	**11.65**	**48.57**	**56.72**	**0.988**	**0.980**	**0.985**
		Varifocal	19.35	92.48	114.82	0.980	0.965	0.980
		Static LF	90.49	426.37	478.87	0.920	0.868	0.910
		Conventional	80.26	112.38	112.10	0.927	0.957	0.983
	Scene c	Ours	**20.94**	**113.78**	**91.60**	**0.979**	**0.963**	**0.938**
		Varifocal	33.43	220.79	271.49	0.968	0.934	0.805
		Static LF	132.23	1193.05	349.55	0.888	0.722	0.792
		Conventional	177.63	279.53	316.82	0.867	0.919	0.773
0.9 dpt uncertainty	Scene a	Ours	**32.03**	**16.14**	**383.53**	**0.982**	**0.992**	**0.962**
		Varifocal	51.34	25.66	687.87	0.974	0.988	0.897
		Static LF	85.64	43.71	657.11	0.937	0.957	0.897
		Conventional	327.50	118.21	714.58	0.871	0.946	0.879
	Scene b	Ours	**14.10**	**55.08**	**87.23**	**0.985**	**0.977**	**0.977**
		Varifocal	26.04	121.69	179.03	0.974	0.953	0.963
		Static LF	92.84	458.69	402.08	0.919	0.856	0.918
		Conventional	82.43	130.62	143.91	0.927	0.949	0.975
	Scene c	Ours	**24.87**	**130.29**	**158.27**	**0.975**	**0.957**	**0.891**
		Varifocal	40.81	292.49	349.69	0.961	0.912	0.747
		Static LF	130.73	1291.82	381.59	0.889	0.702	0.780
		Conventional	178.39	323.95	366.50	0.867	0.904	0.734

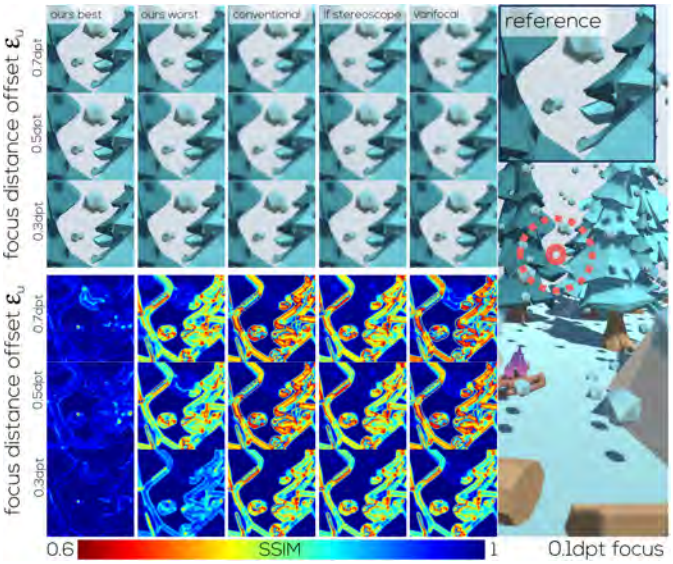

Fig. 9: Qualitative comparison of results in case of 0.1 dpt user focus distance. (right) The ground truth depth-of-field image with the corresponding gaze location and a magnified inset. (left) Simulated perceived images of compared display architectures, in the event of three different focus distance errors. "ours best" shows results for a view volume of size that matches the focus distance error, with a center that is offset from the ground truth focus distance by the amount of the error. "ours worst" depicts the rare case in which the ground truth focus distance is located outside of the view volume, Rows on top show the simulation results, while rows on the bottom show color-coded pixel-wise SSIM values. Please refer to the supplemental material for additional results.

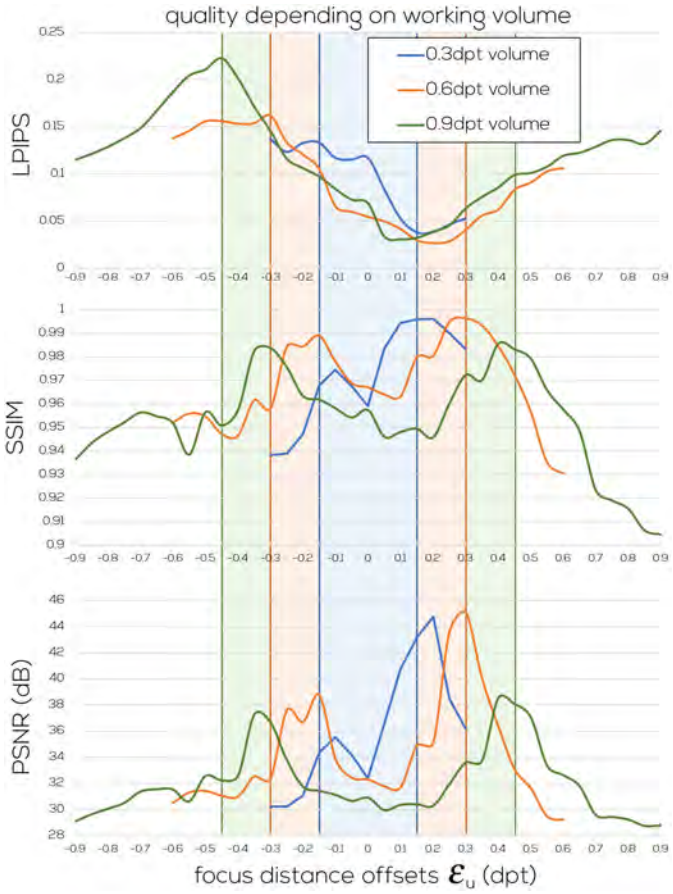

Fig. 10: Example of how the perceived image quality changes depending on view volume and focus distance offset (difference between the center of the view volume and the ground truth focus distance of the user). Results taken from Scene (c) with a ground truth focus distance of 1 dpt.

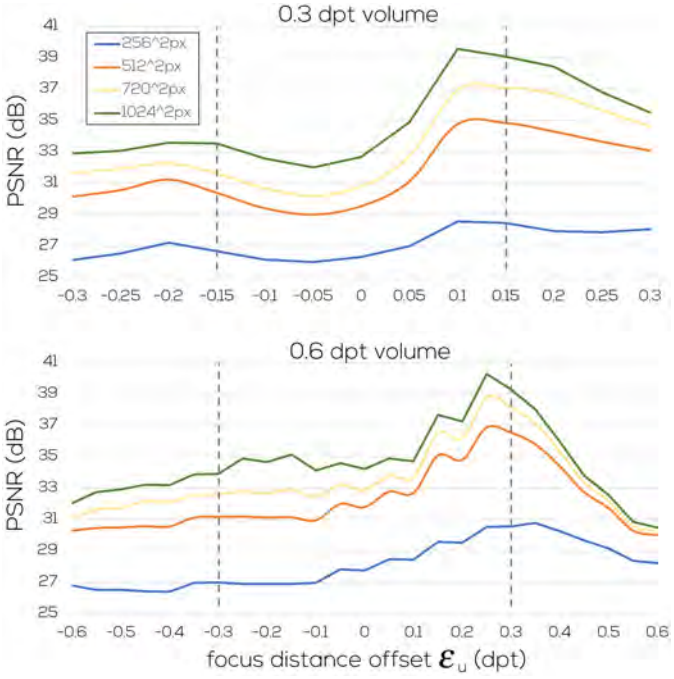

Fig. 11: Effect of front layer resolution on perceived visual quality. We measured the impact of varying front layer resolution on image quality in terms of PSNR for two volume sizes and different focus distance offsets. Similar to Figure 10, the results were acquired with a ground truth focus distance of 1 dpt in Scene (c).

5.4 Impact of front layer resolution

Since the layers in our system are comprised of transmissive LCD panels, the pixel size of the front layer is limited by diffraction (refer to Section 5.1). However, if non-transmissive spatial light modulators, such as LCOS displays, were used, the calculated diffraction limit does not apply any more, and the front layer resolution could be increased. We investigated the impact of various front layer resolutions on the perceived image quality measured in PSNR for two volume sizes and focus distance offsets. As indicated in Figure 11, the perceived image quality correlates with the resolution of the front layer, leading to higher PSNR with increasing resolution.

5.5 Qualitative results

To verify the simulated results presented in Section 5.3, we captured through-the-lens images of our prototype and scene (c), while simulating other display types. For the conventional HMD and the varifocal display, we made the front layer transparent (by setting all pixels to white). For the conventional HMD, we fixed the back layer at 0.5 dpt. For the static light field display, we fixed the layers at 5 dpt and 0.8 dpt respectively. The focus distance of the physical camera was set to a ground-truth focus distance of 3 dpt. The displayed content was created using the inferred focus distance of the user at a focus distance offset of 0.6 dpt. The results are shown in Figure 12. In general, the captures confirm the simulated results. In the best case, our system manages to alleviate the eye tracking error to a high degree, showing a crisp image of the candle in the front. In the worst case, our system still performs better than the other display types. However, the difference in terms of contrast loss compared to the best case is clearly visible.

5.6 Initialization

To quantify the impact of our initialization scheme, we measured the change in decomposition quality with increasing decomposition order m (see Section 3.5). As the order of decomposition increases, the number of convolutions increases, and the performance decreases accordingly. However, higher decomposition orders lead to better quality. Figure 13 shows an example of this relationship using a view volume of ± 0.1 dpt around a user focus distance of 0.2 dpt in scene (b). The image quality in terms of PSNR is shown on the left side and is measured using the average focus distance across the view volume. The corresponding runtime in ms is shown on the right side. In our system, we selected a decomposition order $m = 10$ for the transition mode, which was found to represent the subjectively best compromise between runtime and quality. Overall, the intialization scheme leads to roughly the same quality as the iterative decomposition.

5.7 Runtime

The latency in our system is mainly dependent on the decomposition. Therefore, we evaluated the runtime of a single iteration of the decomposition. The decomposition runtime is correlated to the working volume size (since the size of the circle of confusion increases as well) and to the number of images in the stack. In our implementation, the difference in the focus distance of the focal stack images is 0.1 dpt. Figure 14 shows this dependency for three different resolutions. In our system, we used a resolution of 1440×1440 px (per eye), which corresponds to the maximum resolution that any of our LCD panels can emit while still providing reasonable performance.

6 Conclusions and Future Work

Our results show the impact of the focus estimation confidence on the contrast and the image quality. We believe that our display presents a practical solution to improving the overall contrast of a layered HMD, which is crucial for future immersive VR/AR experiences.

For a given focus distance confidence, the perceived quality and contrast are the lowest when the predicted focus distance perfectly matches the user's actual focus distance, or if the user focus distance is

Fig. 12: Through-the-lens photographs of Scene (c) for various display types. We focused the camera at the ground-truth focus distance of 3 dpt (corresponding to the candle in the front). The eye tracking error is 0.6 dpt. The bottom row shows magnified insets of the capturings. Overall, the through-the-lens images confirm the simulated results. While "ours best" leads to the image with best contrast, "ours worst" still manages to outperform the other display types by a small margin. As expected, the conventional display yields the worst results.

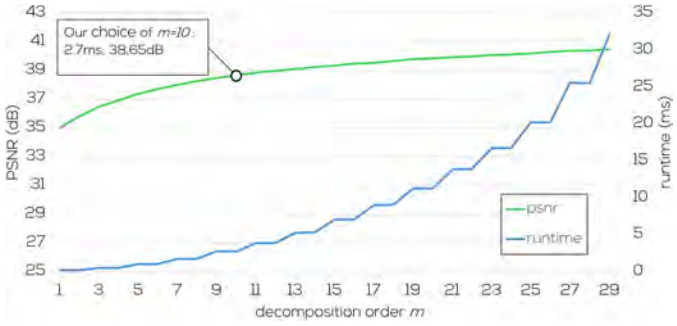

Fig. 13: Image quality in PSNR relative to the decomposition order m of the initialization technique, used in the transition mode for scene (b) and a view volume of 0.2 dpt around the user focus distance. To provide real-time frame rates, we selected $m = 10$, which, in this particular case, yields a PSNR of $38,65$ dB.

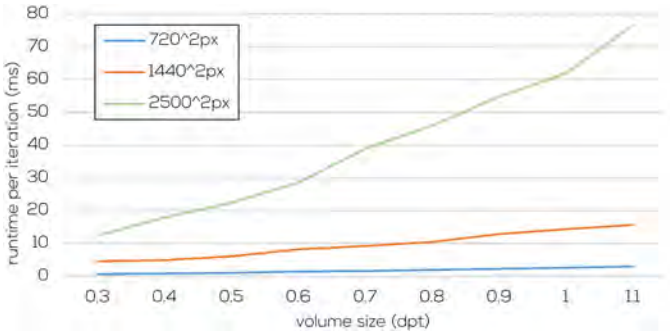

Fig. 14: Runtime of a single iteration of the decomposition with respect to display working volume at various resolutions. As the volume size increases, both the number of images in the focal stack and the size of the circles of confusion on the layers grow, resulting in longer runtimes.

not within the working volume of the display. In the first case ($\varepsilon_u = 0$), the varifocal display would achieve the best results. This issue can be mitigated through additional LCD panels at the cost of brightness, or by spanning a smaller volume and accepting lower contrast when users focus at the outer edges of the confidence interval. However, the aforementioned situation of a perfect match between inferred and actual focus distance represents only a single sample of several possibilities. As shown in Table 1, if we allow for other focus distance offsets, our system outperforms the other display types in terms of quality.

Our eye tracking algorithm incorporates the eye vergence as one of the predictors of the true user focus distance. Prior studies have shown that beyond 0.5 m (2 dpt), vergence becomes an unreliable predictor of focus depth. A variant of our depth-range prediction algorithm could ignore vergence if the predicted depth is beyond this distance, relying only on the predicted gaze point instead. However, vergence can be used as a classifier of different depth ranges that are sufficiently far apart [39], e.g., it could be used to determine if the user focuses between 0.5 and 1 m, 1 m and 2 m, or beyond that. While our solution currently does not account for variable eye tracking accuracy throughout the viewing field, the extension is straightforward.

Instead of using adjustable lenses to shift the virtual image layers in our prototype, we decided to employ a mechanical mechanism. Although adjustable lenses could result in a smaller form factor for the prototype, they also introduce additional aberrations and reduce the field of view, as an additional lens per eye would need to be placed between the layers. This choice would further decrease the supported display volume. We plan to make efforts towards miniaturizing the prototype and increasing its field of view. Furthermore, while our display supports increasing the perceived contrast, its impact on VAC mitigation should be verified in a study with human subjects.

Acknowledgments

This work was supported by Snap, Inc., and the Alexander von Humboldt Foundation funded by the German Federal Ministry of Education and Research.

References

[1] 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*, 36(6):1–13, 2017. doi: 10.1145/3130800.3130892 2

[2] M. S. Arefin, N. Phillips, A. Plopski, J. L. Gabbard, and J. E. Swan. The effect of context switching, focal switching distance, binocular and monocular viewing, and transient focal blur on human performance in optical see-through augmented reality. *IEEE Trans. Vis. Comp. Graph.*, 28(5):2014–2025, 2022. doi: 10.1109/TVCG.2022.3150503 1

[3] M. D. Barrera Machuca and W. Stuerzlinger. The effect of stereo display deficiencies on virtual hand pointing. In *Proc. ACM CHI*, pp. 1–14, 2019. doi: 10.1145/3290605.3300437 1

[4] A. U. Batmaz, M. D. Barrera Machuca, J. Sun, and W. Stuerzlinger. The effect of the vergence-accommodation conflict on virtual hand pointing in immersive displays. In *Proc. ACM CHI*, pp. 1–15, 2022. doi: 10.1145/3491102.3502067 2

[5] P. Chakravarthula, Y. Peng, J. Kollin, H. Fuchs, and F. Heide. Wirtinger holography for near-eye displays. *ACM Transactions on Graphics*, 38(6):1–13, 2019. doi: 10.1145/3355089.3356539 2

[6] J.-H. R. Chang, B. V. K. V. Kumar, and A. C. Sankaranarayanan. Towards multifocal displays with dense focal stacks. *ACM Transactions on Graphics*, 37(6):1–13, 2018. doi: 10.1145/3272127.3275015 2

[7] J.-H. R. Chang, A. Levin, B. V. K. V. Kumar, and A. C. Sankaranarayanan. Towards occlusion-aware multifocal displays. *ACM Transactions on Graphics*, 39(4):68–1, 2020. doi: 10.1145/3386569.3392424 3

[8] A. T. Duchowski, D. H. House, J. Gestring, R. Congdon, L. undefinedwirski, N. A. Dodgson, K. Krejtz, and I. Krejtz. Comparing estimated gaze depth in virtual and physical environments. In *Proc. Symposium on Eye Tracking Research and Applications (ETRA)*, 8 pages, p. 103–110, 2014. doi: 10.1145/2578153.2578168 3

[9] D. Dunn. Required accuracy of gaze tracking for varifocal displays. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 1838–1842. IEEE, 2019. doi: 0.1109/VR.2019.8798273 1, 3, 7

[10] 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 Trans. Vis. Comp. Graph.*, 23(4):1322–1331, 2017. doi: 10.1109/TVCG.2017.2657058 2

[11] C. Ebner, P. Mohr, T. Langlotz, Y. Peng, D. Schmalstieg, G. Wetzstein, and D. Kalkofen. Off-axis layered displays: Hybrid direct-view/near-eye mixed reality with focus cues. *IEEE Trans. Vis. Comp. Graph.*, 29(5):2816–2825, 2023. doi: 10.1109/TVCG.2023.3247077 2

[12] C. Ebner, S. Mori, P. Mohr, Y. Peng, D. Schmalstieg, G. Wetzstein, and D. Kalkofen. Video see-through mixed reality with focus cues. *IEEE Trans. Vis. Comp. Graph.*, 28(5):2256–2266, 2022. doi: 10.1109/TVCG.2022.3150504 1, 2, 4

[13] J. Hensley, T. Scheuermann, G. Coombe, M. Singh, and A. Lastra. Fast summed-area table generation and its applications. *Computer Graphics Forum*, 24(3):547–556, 2005. doi: 10.1111/j.1467-8659.2005.00880.x 5

[14] T. Hirzle, J. Gugenheimer, F. Geiselhart, A. Bulling, and E. Rukzio. A design space for gaze interaction on head-mounted displays. In *Proc. ACM CHI*, pp. 1–12, 2019. doi: 10.1145/3290605.3300855 3

[15] 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 (JOV)*, 8(3):33–33, 2008. doi: 10.1167/8.3.33 1, 2, 3

[16] 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*, 34(4):1–12, 2015. doi: 10.1145/2766922 2, 3, 4, 7, 8

[17] 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):1–36, 2021. doi: 10.1145/3453157 1

[18] 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*, 38(4):99:1–15, 2019. doi: 10.1145/3306346.3322987 2, 6

[19] B. Kress. *Optical Architectures for Augmented-, Virtual-, and Mixed-reality Headsets*. Press Monographs. SPIE Press, 2020. doi: 10.1117/3.2559304 1, 2

[20] D. Lanman and D. Luebke. Near-eye light field displays. *ACM Transactions on Graphics*, 32(6):1–10, 2013. doi: 10.1145/2508363.2508366 2

[21] S. Lee and H. Hua. A robust camera-based method for optical distortion calibration of head-mounted displays. *Journal of Display Technology*, 11(10):845–853, 2014. doi: 10.1109/JDT.2014.2386216 6

[22] Y. Lee, C. Shin, A. Plopski, Y. Itoh, T. Piumsomboon, A. Dey, G. Lee, S. Kim, and M. Billinghurst. Estimating gaze depth using multi-layer perceptron. In *2017 International Symposium on Ubiquitous Virtual Reality (ISUVR)*, pp. 26–29. IEEE, 2017. doi: 10.1109/ISUVR.2017.13 3

[23] S. Liu, D. Cheng, and H. Hua. An optical see-through head mounted display with addressable focal planes. In *Proc. IEEE ISMAR*, pp. 33–42, 2008. doi: 10.1109/ISMAR.2008.4637321 2

[24] K. J. MacKenzie, R. A. Dickson, and S. J. Watt. Vergence and accommodation to multiple-image-plane stereoscopic displays:"real world" responses with practical image-plane separations? *SPIE Journal of Electronic Imaging (JEI)*, 21(1):011002–011002, 2012. doi: 10.1117/12.872503 2

[25] A. Maimone and H. Fuchs. Computational augmented reality eyeglasses. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 29–38. IEEE, 2013. doi: 10.1109/ISMAR.2013.6671761 2

[26] K. Maruyama, K. Takahashi, and T. Fujii. Comparison of layer operations and optimization methods for light field display. *IEEE Access*, 8:38767–38775, 2020. doi: 10.1109/ACCESS.2020.2975209 1

[27] I. Matsumura, S. Maruyama, Y. Ishikawa, R. Hirano, K. Kobayashi, and Y. Kohayakawa. The design of an open view autorefractometer. In *Advances in Diagnostic Visual Optics: Proceedings of the Second International Symposium, Tucson, Arizona, October 23–25, 1982*, pp. 36–42. Springer, 1983. doi: 10.1007/978-3-540-38823-4_5 3

[28] 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*, 36(6):237–1, 2017. doi: 10.1145/3130800.3130846 2, 5

[29] 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*, 34(4):59, 2015. 2

[30] N. Padmanaban, R. Konrad, T. Stramer, E. A. Cooper, and G. Wetzstein. Optimizing virtual reality for all users through gaze-contingent and adaptive focus displays. *Proceedings of the National Academy of Sciences*, 114(9):2183–2188, 2017. doi: 10.1073/pnas.1617251114 2

[31] N. Padmanaban, R. Konrad, and G. Wetzstein. Autofocals: Evaluating gaze-contingent eyeglasses for presbyopes. *Science advances*, 5(6):eaav6187, 2019. doi: 10.1126/sciadv.aav6187 3

[32] Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein. Neural holography with camera-in-the-loop training. *ACM Transactions on Graphics*, 39(6):1–14, 2020. doi: 10.1145/3414685.3417802 2

[33] D. Pi, J. Liu, and Y. Wang. Review of computer-generated hologram algorithms for color dynamic holographic three-dimensional display. *Light: Science & Applications*, 11(1):231, 2022. doi: 10.1038/s41377-022-00916-3 2

[34] K. Rathinavel, H. Wang, A. Blate, and H. Fuchs. An extended depth-at-field volumetric near-eye augmented reality display. *IEEE Trans. Vis. Comp. Graph.*, 24(11):2857–2866, 2018. doi: 10.1109/TVCG.2018.2868570 2

[35] K. Rathinavel, G. Wetzstein, and H. Fuchs. Varifocal occlusion-capable optical see-through augmented reality display based on focus-tunable optics. *IEEE Trans. Vis. Comp. Graph.*, 25(11):3125–3134, 2019. doi: 10.1109/TVCG.2019.2933120 2

[36] D. Schmalstieg and T. Höllerer. *Augmented Reality - Principles and Practice*. Addison-Wesley Professional, June 2016. 1

[37] B. Schwerdtfeger, R. Reif, W. A. Gunthner, G. Klinker, D. Hamacher, L. Schega, I. Bockelmann, F. Doil, and J. Tumler. Pick-by-vision: A first stress test. In *Proc. IEEE ISMAR*, pp. 115–124. IEEE, 2009. doi: 10.1109/ISMAR.2009.5336484 1

[38] 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. doi: 10.1109/TIP.2018.2839263 4, 5

[39] T. Toyama, J. Orlosky, D. Sonntag, and K. Kiyokawa. A natural interface for multi-focal plane head mounted displays using 3d gaze. In *Proc. International Working Conference on Advanced Visual Interfaces (AVI)*, pp. 25–32, 2014. doi: 10.1145/2598153.2598154 10

[40] M. Weier, T. Roth, A. Hinkenjann, and P. Slusallek. Predicting the gaze depth in head-mounted displays using multiple feature regression. In *Proc. Symposium on Eye Tracking Research and Applications (ETRA)*, pp. 1–9, 2018. doi: 10.1145/3204493.3204547 3

[41] 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*, 30(4):1–12, 2011. doi: 10.1145/2010324.1964990 2, 3

[42] G. Wetzstein, D. Lanman, M. Hirsch, and R. Raskar. Tensor Displays: Compressive Light Field Synthesis using Multilayer Displays with Directional Backlighting. *ACM Transactions on Graphics*, 31(4):1–11, 2012. doi: 10.1145/2185520.2185576 3, 5

[43] A. Wilson and H. Hua. High-resolution optical see-through vari-focal-plane head-mounted display using freeform alvarez lenses. In *Digital Optics for Immersive Displays*, vol. 10676, pp. 384–390. SPIE, 2018. doi: 10.1117/12.2315771 2

[44] W. Wu, P. Llull, I. Tosic, N. Bedard, K. Berkner, and N. Balram. Content-adaptive focus configuration for near-eye multi-focal displays. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016. doi: 10.1109/ICME.2016.7552965 2

[45] 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)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068 8