# A Perception-driven Hybrid Decomposition for Multi-layer Accommodative Displays

Hyeonseung Yu, Mojtaba Bemana, Marek Wernikowski, Michał Chwesiuk, Okan Tarhan Tursun, Gurprit Singh, Karol Myszkowski, Radosław Mantiuk, Hans-Peter Seidel, Piotr Didyk

**Abstract**— Multi-focal plane and multi-layered light-field displays are promising solutions for addressing all visual cues observed in the real world. Unfortunately, these devices usually require expensive optimizations to compute a suitable decomposition of the input light field or focal stack to drive individual display layers. Although these methods provide near-correct image reconstruction, a significant computational cost prevents real-time applications. A simple alternative is a linear blending strategy which decomposes a single 2D image using depth information. This method provides real-time performance, but it generates inaccurate results at occlusion boundaries and on glossy surfaces. This paper proposes a perception-based hybrid decomposition technique which combines the advantages of the above strategies and achieves both real-time performance and high-fidelity results. The fundamental idea is to apply expensive optimizations only in regions where it is perceptually superior, e.g., depth discontinuities at the fovea, and fall back to less costly linear blending otherwise. We present a complete, perception-informed analysis and model that locally determine which of the two strategies should be applied. The prediction is later utilized by our new synthesis method which performs the image decomposition. The results are analyzed and validated in user experiments on a custom multi-plane display.

**Index Terms**—3D displays, Rendering, Accommodation, Perception

## 1 Introduction

In recent years, head-mounted displays (HMDs) have emerged as a major virtual (VR) and augmented reality (AR) technology and currently they have many potential applications in a diverse set of fields including gaming, video, medicine, simulation and aviation. Stereo HMDs can display 3D content with binocular disparity, which is one of the critical cues for stereopsis and depth perception of the brain. As the use of binocular disparity in HMDs has already been successfully commercialized, research efforts are recently getting directed towards enhancing 3D perception by introducing a support for other types of cues. A critical requirement for a faithful reconstruction of virtual 3D content is the reproduction of correct accommodation cues, which allows a natural depth perception by triggering changes in the focal distance of the eye [5, 21]. However, developing HMDs with correct accommodation cues is an extremely challenging task due to the limitations imposed by optics on the hardware design. Any improvement in this direction must satisfy the requirements from a consumer product such as having a small form factor but usually there is a trade-off between these requirements and optical capabilities of the display such as the field of view (FOV) and display resolution [13]. In addition to these hardware challenges, generating 3D content for such displays is another important issue since it requires efficient processing of a larger amount of data compared to 2D images. Furthermore, there is always a concern of having compatibility with different display architectures [19].

Recent studies have shown that multi-layer displays, such as multi-plane displays or light-field displays, are practical solutions for HMDs to provide near-correct accommodation cues [13, 19]. A crucial step of rendering in a multi-layered system is the decomposition of an input scene into layers for a correct 3D perception [37]. The most straightforward decomposition method is linear blending (LB), where the input is a single viewpoint image with a depth map [2, 29]. Although this technique is computationally efficient, it usually fails at occlusion boundaries or non-Lambertian surfaces. To overcome this limitation, two approaches have been proposed: retinal optimization (RO) [35, 37] and light-field synthesis (LFS) [16, 25], which optimize the decomposition based on a focal stack and a 4D light field, respectively. The improved quality comes at a high computational cost of the optimization (5 Hz at 512×512 resolution as reported by Mercier et al. [35]) and input generation. In addition, although these techniques perform better at occlusion boundaries [48], they may perform worse in driving the eye accommodation [35].

In order to combine the desired features of different algorithms, the most promising solution would be designing a hybrid decomposition technique. Such an approach could select the decomposition method locally depending on the scene content in order to obtain the best perceptual quality possible. For real-time rendering applications, this type of hybrid decomposition has to be implemented efficiently. Thanks to the recent developments in GPU hardware, new cards introduce separate cores for massively computational tasks (e.g. recently announced Nvidia RTX platform) which encourages such content-dependent local optimizations to be performed in parallel to the traditional graphics pipeline. However, in order to propose a robust hybrid algorithm, a clear understanding of the perceptual quality differences among various decomposition methods is required. So far, there has been very little research comparing the visual quality of LB, LFS and RO methods. In addition, the conditions which lead to failure of LB method at occlusion boundaries are not thoroughly investigated in previous works.

To address these issues, we provide a perceptual evaluation of different decomposition methods and propose a perception-driven hybrid decomposition technique. In the first part of our paper, as a preliminary step towards the hybrid decomposition, we introduce an improved gaze-contingent LFS method that generates the input viewpoints exclusively inside the pupil. We demonstrate that this solution achieves similar results to RO but with a significantly lower amount of computational cost. Consequently, we focus on the gaze-contingent LFS, and omit RO in our considerations. In the second part, we propose a perceptual evaluation methodology to determine for which multi-plane display configurations and scene content the inexpensive LB can be applied without a loss of visual quality and when the gaze-contingent LFS is necessary. In our analysis, we focus on texture and occlusion


- *Hyeonseung Yu, Mojtaba Bemana, Okan Tarhan Tursun, Gurprit Singh, Karol Myszkowski, Hans-Peter Seidel are with Max-Planck Institute for Informatics. E-mail: hyu@mpi-inf.mpg.de, mbemana@mpi-inf.mpg.de, okan.tursun@mpi-inf.mpg.de, gsingh@mpi-inf.mpg.de, karol@mpi-inf.mpg.de, hpseidel@mpi-inf.mpg.de.*
- *Marek Wernikowski, Michał Chwesiuk and Radoslaw Mantiuk are with West Pomeranian University of Technology. E-mail: mwernikowski@wi.zut.edu.pl, mchwesiuk@wi.zut.edu.pl, rmantiuk@wi.zut.edu.pl.*
- *Piotr Didyk is with Università della Svizzera italiana. E-mail: piotr.didyk@usi.ch.*


*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*

boundaries which are important for driving accommodation [29, 33] and depth order perception [48]. We derive the detection thresholds through a series of perceptual experiments, which allows us to establish the selection rule for the decomposition algorithm such that:

1. when LB and LFS are visually indistinguishable, we select the LB, and
2. when LB and LFS are distinguishable, we choose the method which yields results closer to the ground truth.

Based on the selection rule, we describe the hybrid decomposition approach which combines linear blending and light-field synthesis methods. To further improve the performance, we also take the foveal and peripheral vision characteristics into account. Consequently, we propose a content-dependent and gaze-contingent hybrid decomposition algorithm for multi-layered accommodative displays, which enables real-time rendering performance and high-quality reconstruction.

The main contributions of this work are:

- a gaze-dependent viewpoint sampling of LFS for enhanced reconstruction quality,
- a series of targeted perceptual experiments that measure the differences in the visual quality obtained by LB and LFS for various spatial frequencies, luminance contrasts, depth configurations, and eccentricities,
- a domain-specific structural similarity index (SSIM) calibration for visible difference prediction between the LB and LFS that generalizes perceptual insights beyond the scope of the perceptual experiments,
- a unified optimization framework for the LB and LFS decompositions,
- an efficient adaptation of the simultaneous algebraic reconstruction technique (SART) to CUDA for the real-time decomposition.

## 2 Background and Previous Work

In this section, we give an overview of near-eye displays supporting accommodation cues, image decomposition algorithms targeted for such displays, as well as selected aspects of foveated rendering that are central for this work.

### 2.1 Accommodative near-eye displays

Multi-plane displays Multi-plane displays project images on different depth planes and form near-correct 3D volumetric images. The system architecture can be classified into two categories: systems based on time-multiplexing with switchable lenses [12, 28] or systems based on beam splitters and multiple physical screens [2, 29]. Time-multiplexing systems can be designed in smaller form factors, but the requirement for high-refresh-rate screens and fast tunable-focus devices is a major obstacle. Although multi-screen systems have a major drawback in its large form factor, they offer a larger FOV than time-multiplexing systems. Therefore, we employ this design to test our rendering strategy. Another major obstacle of both architectures is the requirement for eye tracking since the images are generated for a fixed viewing position. Recently, focal surface displays have been developed to represent continuous 3D imagery [34]. They eliminated the need for eye tracking in the case of single plane generation, but they are computationally demanding and based on expensive LCoS SLMs. Another approach to avoid the eye-tracking is to perform per-region optimization at multiple gaze points, but it requires costly optimization and precise calibration of the eye rotation axis [24]. Since eye-tracking is an essential component in practical multi-plane system settings, we exploit the eye-tracking system further to develop foveated rendering strategy.

Light-field displays The light-field display controls the 4D ray space of the light generated by the display to produce the motion parallax and vergence cues. Recently, light-field displays supporting focus cues have been proposed based on microlenses [14, 22]. However, those designs have an intrinsic trade-off between the angular and spatial resolution. Light-field displays based on multi-layered architecture [15, 30, 36] have been demonstrated as an efficient platform for providing focus cues. Our rendering strategy is mainly built on the principle of additive light-field displays with accommodation cues [36].

Other methods Holographic displays can project a replica of real-world scenes and provide accurate focus cues [47]. However, the limited pixel size and resolution of digital wavefront modulators impose a significant trade-off between the eyebox size and FOV [31]. Another approach is to change the depth of 2D image plane dynamically with focus-tunable devices [3, 8]. Although viewers can observe the images with correct accommodation cues, the requirement for a dynamic system may lead to latency issues. Instead of generating complete focus cues, the vergence-accommodation conflict also can be alleviated by projecting all-in-focus images [17]. However, this method has a trade-off between the spatial resolution and the reproducible focus range. Recently, it is also demonstrated that proper rendering of chromatic aberration can effectively trigger accommodation without changing optical focus cues [6].

### 2.2 Decomposition algorithms for multi-layered displays

Light-field displays In multi-layered light-field displays, the light fields are parametrized by a group of pixels on multiple layers. For multiplicative displays, the optimization system is described in tensor form and solved by various factorization algorithms [16, 45]. Additive light-field displays based on the polarization LCDs [23] or incoherent summation of pixel intensities reflected from holographic optical elements [24] have also been proposed. For those architectures, LFS is formulated with a linear least-squares error problem and solved with the simultaneous algebraic reconstruction technique (SART) for online calculation [4] or the trust-region method [7] for offline calculation. In LFS, generation of target light fields requires high computational cost, and real-time performance is only possible by reducing the number of iterations [16, 23]. To enhance the rendering speed, an adaptive sampling strategy was proposed [11], but the performance improvement was only demonstrated for offline rendering scenarios. Our method saves the computational cost both in generating the target light fields and in decomposition through selective rendering and optimization. Furthermore, the modified formulation of the SART adapted to CUDA enables us to achieve the real-time rendering with good quality.

Multi-plane displays In multi-plane displays, the linear blending rule assigns pixel values proportional to the distance between a target point and display planes [2]. Although it can effectively trigger accommodation [29], occlusion boundaries and non-Lambertian surfaces are imperfectly rendered in LB due to the simple consideration of a single image and depth map. In order to correctly generate artifact-free scenes, the retinal optimization (RO) [35, 37] which optimizes a focal stack has been proposed. However, the target focal stack in fact implicitly contains the 4D light-field information [27]. Therefore, LFS which optimizes the 4D light fields also can be employed in multi-plane display architecture. We provide a short mathematical derivation of the RO in the context of LFS in Supplementary Information A. Our method is based on LFS since the implementations of current LFS algorithms are demonstrated to be more efficient than RO. We also revisit LFS in the context of gaze-contingent rendering for improving the perceived image quality and reducing the computational cost.

### 2.3 Foveated rendering

Foveated rendering uses gaze information to improve rendering efficiency by reducing quality for the periphery. This is usually achieved by reducing the density of rendered image samples with increasing eccentricity [10, 38, 41]. In accommodative light field displays, Sun et al. [40] propose a foveated rendering solution, which accounts for depth information and the current state of the accommodation to choose optimal ray directions in the OptiX renderer. In our work, the ray selection is dictated by the choice of local decomposition technique for multi-plane displays and supported by an analysis of local luminance contrast and visibility of artifacts caused by the LB.

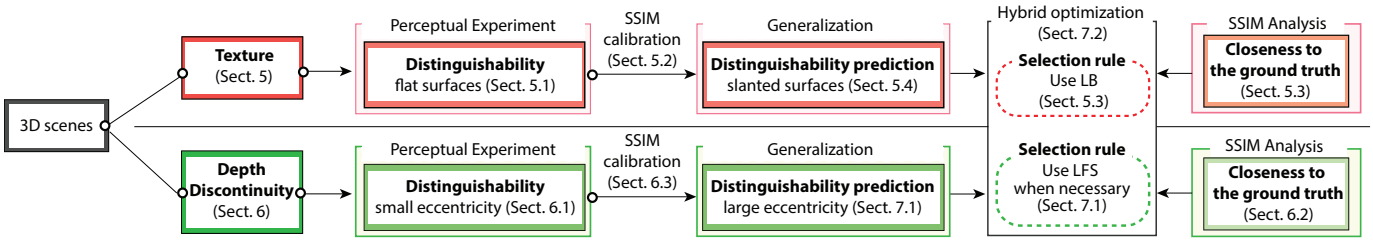



Fig. 1: Overview of the methodology in our work. We study 3D scenes based on texture and depth discontinuity. The limited cases are investigated with perceptual experiments and extended to the general cases through the prediction using custom calibrated SSIM. We derive the selection rules from the predicted distinguishablity between LB and LFS and closeness to the ground truth obtained by SSIM analyses. Then, we finally develop the hybrid optimization framework based on the selection rules.

## 3 Overview

We develop the perceptual evaluation methods and the hybrid decomposition of a gaze-contingent LFS (Sect. 4) and LB. Our analysis pipeline is outlined in Fig. 1. We analyze the 3D scenes based on texture (Sect. 5) and depth discontinuity (Sect. 6), which are important for quality perception and driving accommodation. Ideally, we aim to test all possible scenarios through the perceptual experiments. However, since the parametric space of texture and depth discontinuity is vast, we perform the perceptual experiments on distinguishability between LFS and LB in a limited parametric space. To explore the full space, we calibrate a visual quality metric SSIM [42] to predict the experimental outcomes and predict the distinguishability in general cases. Our employment of SSIM is motivated by a recent study showing that advanced metrics such as SSIM and HDR-VDP [32] provide a similar and good prediction on a narrow, well-defined task after proper training and calibration with relevant perceptual data [1]. Specifically, the perceptual experiments are conducted for flat textured surfaces in Sect. 5.1 and depth discontinuities at small eccentricities in Sect. 6.1. Through the calibrations of SSIM in Sect. 5.2 and Sect. 6.3, we predict the distinguishability in general cases such as slanted textured surfaces in Sect. 5.4 or depth discontinuities at large eccentricities in Sect. 7.1. For selecting a proper algorithm when LFS and LB are distinguishable, we perform SSIM analysis to find the algorithm closer to the ground truth in Sect. 5.3 and Sect. 6.2. Finally, we choose the best decomposition algorithm, which is LB for textured surfaces and LFS for depth discontinuities depending on depth difference, luminance contrast and eccentricities. Since the transition between LFS and LB is required at depth discontinuity, we develop the selection rule in Sect. 7.1 and propose the hybrid optimization framework in Sect. 7.2.

## 4 Gaze-contingent Light Field Synthesis

For our hybrid decomposition strategy, we evaluated the existing decomposition methods for multi-layer displays with respect to computational complexity and visual quality criteria. LB is a fast decomposition method, and it is suitable for the regions where an accurate reconstruction is not required. On the other hand, when a high quality reconstruction is required, the hybrid decomposition algorithm should select more complex methods such as LFS and RO. While LFS reconstructs a sparse set of lightfield views, RO reproduces a focal stack rendered from dense light fields inside the pupil. Although LFS is computationally more efficient than RO, a recent study shows that LFS suffers from contrast degradation and RO might be a better alternative for preserving the contrast [24]. However, our insight is that the loss of contrast in LFS mainly originates from using a wide eye box which is larger than the pupil size, where some of the viewpoints fall outside observer's pupil [16, 24]. On the contrary, RO provides a higher level of contrast by rendering the dense light fields exclusively inside the pupil and further processing them to generate focal images at multiple depths.

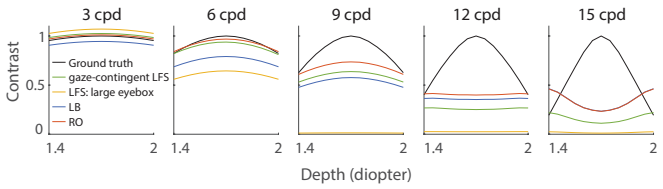



Fig. 2: Contrast curves for various optimization algorithms for various spatial frequencies. While LFS with large eye box exhibits significant contrast reduction for high spatial frequencies, gaze-contingent LFS shows much higher contrast values over the entire frequency range, providing similar quality to LB or RO.

Both LFS and RO might be used to produce high quality outputs when required by a hybrid decomposition algorithm. But the issue of contrast degradation has to be addressed to get the benefit of LFS. To this end, we propose a gaze-contingent viewpoint sampling approach to enhance LFS image quality compared to the implementation using a wide eye box. Our approach is to generate light-field viewpoints only inside the pupil, using the pupil position from an eye tracker. This solution effectively avoids the contrast degradation in LFS method. The gaze-contingent method requires the addition of an eye tracker device to the hardware but as we discussed in Sect. 2, this requirement applies to any practical decomposition method for multi-focal displays.

We validate the quality of the proposed gaze-contingent LFS method using simulated contrast curves of the reconstructed images from various decompositions (Fig. 2). The contrast curves show the magnitude of luminance contrast for different spatial frequencies with respect to accommodation depth. In accommodative displays, the contrast of the images should be maximized at the object plane because a higher gradient of the contrast curve more effectively drives the accommodation toward the object plane [39]. In order to obtain the contrast curves, we first generate retinal images at various focal depths between 1.4 D and 2 D. We chose the 0.6 D gap since it is widely used to attain sufficient resolution for triggering accommodation at intermediate planes and minimizing the number of display planes [29]. Then we use Fourier transform to extract the luminance values at the target spatial frequency. Finally, we normalize all values with the peak value of the contrast curve of the ground truth. A similar analysis has been performed by Lee [24]. We analyze the ground truth, gaze-contingent LFS, LFS with large eye box, LB and RO. During our evaluations, we set the number viewpoints to 13 inside a 4 mm-diameter pupil for gaze-contingent LFS. We empirically found that using larger number of views does not improve the image quality for gaze-contingent LFS. The large eye box case assumes $5 \times 5$ viewpoints inside an $8 \times 8$ mm eye box. The sinusoidal patterns of various spatial frequencies are projected in the middle plane between two display layers placed at 1.4 D and 2 D, respectively. The resolution is set to 15 cpd, which is the maximum resolution supported by our display. The analysis shows that LFS with a large eye box significantly degrades the quality beyond approximately 6 cpd. In contrast, the gaze-contingent LFS provides a quality comparable to RO or LB for 3–9 cpd, which is the critical range for driving accommodation [29, 33]. The noticeable deviations are observed for high spatial frequencies; however, all algorithms already fail to reproduce the correct contrast curve due to the limited frequency support of the display. The maximum reproducible frequency increases as the display separation decreases [37]. Therefore, we can conclude

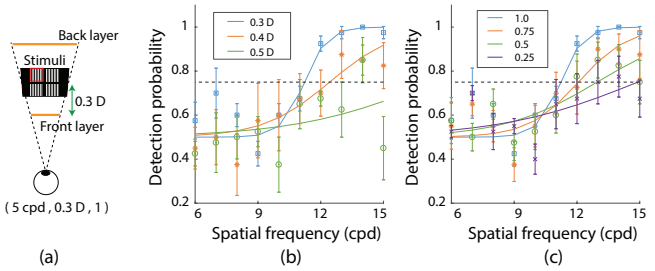



Fig. 3: The configuration of displays and stimuli (a). The probability of detecting the difference between LB and LFS methods for various depth (b) and contrast levels (c).

that the gaze-contingent LFS attains the quality offered by RO and is suitable for using with LB in a hybrid decomposition approach. From now, we refer to gaze-contingent LFS simply as LFS.

## 5 Effect of Texture on Decomposition

LB method performs poorly in regions which are affected by occlusion [37]. However, it preserves the contrast relatively well in other regions (see Fig. 2). Therefore, the use of LB still can be a good option on textured regions except at those problematic regions around occlusion boundaries where LFS gives a better result than LB. The previous work shows that the difference between the two methods is noticeable when the content has high spatial frequencies [37]. Motivated by this observation, we aim to find the spatial frequency threshold where we should switch from one decomposition method to the other in order to get the optimal quality. Our analysis includes showing both flat and slanted surfaces with various slopes. We first conduct perceptual experiment to investigate the conditions when two algorithms are indistinguishable for an observer viewing flat surfaces on our prototype display. Then, we perform an analysis using an objective quality metric in order to generalize our findings to slanted surfaces. In addition to allowing the evaluation of different scene configurations, use of an objective metric helps avoiding any issues due to the lack of ground truth as well. Unfortunately, there is not a domain-specific metric designed for such an evaluation. Therefore, we take an existing full-reference quality metric and calibrate it using the data obtained from our perceptual experiments. For this purpose, we use the Structural Similarity (SSIM) metric which is widely used for the objective evaluation of visual quality in other domains [42].

### 5.1 Perceptual experiment

We conducted the perceptual experiment on a two-plane prototype display in a monocular viewing setting. The stimuli consists of two pairs of flat sinusoidal patterns. One pair contains two identical patterns generated using only LFS, and the other pair contains two different patterns generated using LFS and LB. In the experiment, we used two-alternative forced choice (2AFC) procedure and asked the participants to select the pair of patterns which looks different from each other. While two pairs are shown at the top and bottom positions, the order of patterns is completely randomized among trials. We computed the probability of detection from the number of correct responses for different combinations of Michelson contrast, stimuli depth and spatial frequencies from 6 to 15 cpd. Fig. 3(a) shows representative stimuli used in our experiment, where the LB stimulus has a red frame around the pattern. For that stimuli, the correct response is the top pair. In total, five participants took the experiment. All participants were naïve, paid, and have normal or corrected-to-normal vision. The display resolution is 15 cpd and the display separation is set to 0.6 D. Please see Sect. 8.2 for more details on the experimental setup.

We take the frequency which corresponds to 75% detection probability as the detection threshold, which is computed by fitting a psychometric sigmoidal function to the collected data. The detection probabilities from the experiment and fitted sigmoids are shown in Fig. 3 (b-c). Fig. 3 (b) is obtained for various depths of the stimuli, while the Michelson contrast is fixed at 1. The depth is measured as the distance from the front display, where 0.3 D corresponds to the middle plane. The frequency threshold has the smallest value for the middle plane stimuli, where the reconstruction quality of decomposition algorithms is the lowest [29, 37]. Fig. 3 (c) shows the results for various contrasts, while the stimuli depth is fixed to 0.3 D. These results indicate that the frequency threshold increases as the contrast decreases.

### 5.2 Calibrating SSIM

The above experiment considers only a small subset of different texture and depth configurations which can occur in complex scenes. One option to investigate a wider range of stimuli is performing more extensive perceptual experiments. Instead, in this work, we propose to rely on image quality metrics which have been recently demonstrated to be successful in simulating visibility of different artifacts when calibrated on a problem-specific dataset [1, 46]. Consequently, we adapt SSIM metric for predicting distinguishability between LFS and LB methods and use it in further investigation. An additional and critical benefit of such a strategy is that it also allows us to compare the decomposition techniques to ground-truth images. This is challenging using perceptual experiments due to the lack of a reference light-field display.

Our SSIM-based metric takes as an input two perceived images, simulated for a specific focus, and computes an SSIM map. The metric later takes the minimum value of the map as the dissimilarity measure between the two images. We first use this procedure to simulate the previous experiment. To this end, we computed the dissimilarity index between LFS and LB methods for different combinations of luminance contrast, frequency, and depth, assuming that the observer focuses on the target object plane. Fig. 4 (a) and (b) show the results for the stimuli when the Michelson contrast of the stimuli is fixed to 1 and 0.5. Smaller values of the maps indicate a larger difference between the results of the two methods. Similar to the result in Fig. 3 (b), the transition behavior is observed around 12 cpd for the middle plane (0.3 D), and the transition point moves towards higher frequencies for stimuli closer to the display plane. Fig. 4 (b) reveals that the SSIM values overall increase for lower luminance contrast, which is in agreement with Fig. 3 (c).

To use the metric as a visibility predictor, we seek an SSIM threshold such that it corresponds to the visibility threshold. In other words, all the image regions for which the SSIM index is above the SSIM threshold should contain only invisible differences while the regions with smaller SSIM values contain visible differences. We determine the optimal SSIM threshold as the value which minimizes the RMS error between the predicted and measured frequency thresholds obtained in the experiment in Sect. 5.1 . The lowest error was obtained for the SSIM threshold 0.9. The rest of our evaluations are based on this value.

### 5.3 Comparison with ground truth

We apply SSIM analysis to select the algorithm which is closer to the simulated ground truth. We first obtain two SSIM maps in the same way as Fig. 4 (a) by comparing LFS with ground truth, and LB with ground truth. Then we take the pixel-wise difference between the SSIM map of LFS and LB. The result is shown in Fig. 4 (d). We observe that LB is better at reproducing the ground truth in the region inside the dashed half-circle, where LFS and LB are distinguishable according to our previous analysis. Outside this region, LFS performs better particularly at low frequencies, but it is still acceptable to use LB due to indistinguishability.

Interestingly, these results suggest that the computationally efficient LB provides higher fidelity reconstruction compared to the computationally expensive LFS on high spatial frequencies. Although previous study [37] and our analysis (Fig. 2) suggest that such high contrast reconstruction can lead to incorrect contrast curve, the eye accommodation is dominantly driven by 4-8 cpd and the failures of LB in reproducing contrast gradient at high frequencies are negligible [29]. Furthermore, it should be noted that even LFS or RO fail to reproduce the correct contrast curves in such cases, as shown in Fig. 2. Therefore, we choose LB as the best algorithm which provides high contrast in retinal images.

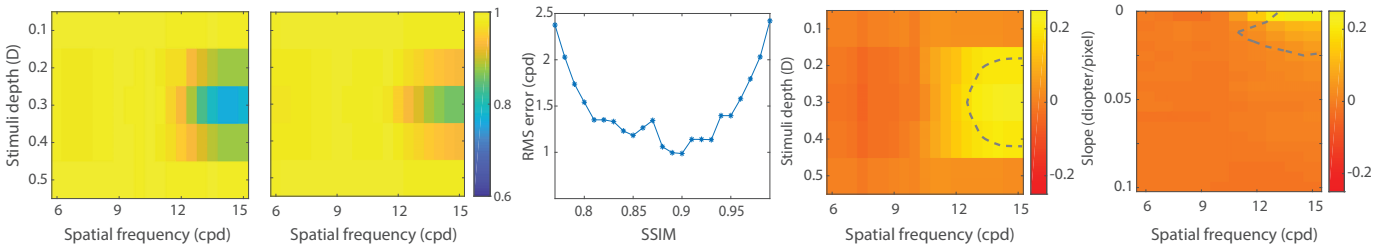



Fig. 4: The minimum SSIM values of the SSIM maps between the LB and LFS for (a) contrast = 1 and (b) contrast = 0.5. (c) The RMS error between the predicted cutoff frequencies and experimental results. (d) The SSIM against the ground truth. The positive region: LB is closer to the ground truth. The negative region: LFS is closer to the ground truth. The yellow region bounded by the dashed line represents the conditions where two methods are distinguishable. (e) The ground truth comparison for slanted surfaces.

### 5.4 Generalization to slanted surfaces

In many studies, the quality of reconstruction has been tested on planar surfaces at a fixed depth [24, 29, 37]. However, most 3D scenes contain various slanted surfaces. Here, we extend our analysis to slanted surfaces with various slopes. At each spatial frequency, we generate slanted surfaces up to the maximum slope of 0.1 D/pixel. In our display prototype with the 0.6 D separation, this maximum slope corresponds to a 6-pixel-wide slanted surface extending from the front display to the back display. Since a fewer number of pixels cannot fully represent one cycle of the minimum spatial frequency, we regard steeper surfaces as occlusion boundaries. In the previous analysis of flat surfaces, we compared the focal images at the target stimulus plane. In the presence of a slanted surface, however, the reconstruction quality should be checked at every possible focal state. Therefore, we first compute 7 focal images between two layers with a step size of 0.1 D. For each focal image of each algorithm, we find the minimum SSIM value in the comparison against the ground truth focal image. Among all focal depths, we again select the minimum SSIM to find the worst case. Then we take difference between the SSIM map of LFS and LB to compute the closeness to the ground truth, as shown in Fig. 4 (e). The border of the distinguishable region is indicated with the gray dashed line. Similar to the flat surfaces, two methods are distinguishable for high spatial frequency texture at low slopes. Inside this distinguishable region, LB still performs better than LFS.

In summary, our analyses reveal that for flat and slanted surfaces with sinusoidal patterns as textures, LB and LFS methods are distinguishable only for high spatial frequency textures, and LB provides the higher fidelity reconstruction when they are distinguishable. Since this holds for foveal vision, it is evident that the same algorithm holds for peripheral vision because contrast sensitivity declines in the peripheral region.

## 6 Effect of Depth Discontinuity on Decomposition

Another factor which affects the decomposition quality is the depth difference between two surfaces with an occlusion boundary. Here, we investigate the distinguishability between LFS and LB as a function of depth difference, luminance contrast, and eccentricity. Contrary to the analysis on spatial frequency in Sect. 5, we found that LFS is always closer to the ground truth compared to LB, but we are aiming to clearly identify the conditions in which LB can still be employed without causing any visible loss of quality.

### 6.1 Perceptual experiment

Similar to the perceptual experiment in Sect. 5.1, we follow 2AFC procedure and ask the participants to select the pair consisting of different patterns. For each luminance contrast and eccentricity, we employ the Quest procedure to find the threshold of depth difference at which LB becomes distinguishable [44]. The depth difference ranges from 0.05 D to 0.6 D with a 0.05 step size. Two representative stimuli are illustrated in Fig. 5 (a). While the foreground objects are fixed on the front display, we change the depth of occluded objects. For the experiments at higher eccentricities, the gaze direction is guided by a target green cross and observers' gaze position is monitored using the eye tracker. In order to avoid incorrect measurements due to accidental glances, the stimulus is hidden when the gaze position slightly deviates from the target cross. The whole set of stimuli spans 3° of visual angles. In order to avoid image degradation due to the aberration near the boundaries of the display, we fix the position of the stimuli at the center of the display and change the position of target cross to control stimulus eccentricity.

The results of this experiment are shown in Fig. 5 (b–d). We observe an increase in depth difference thresholds with respect to eccentricity as expected. This implies that the human visual system (HVS) is less sensitive to the incorrect edges generated by LB in the peripheral visual field and it provides us the flexibility of using LB instead of LFS at edges located in the periphery to improve the performance. Another observation is that the difference between LB and LFS decompositions is highly distinguishable at low luminance contrast edges, which is a finding that is in the opposite direction of the analysis on texture, where the difference between LB and LFS is reduced with the luminance contrast reduction as shown in Fig. 4 (a,b). Notice that at the occlusion boundaries, the mixed signals from the focused and defocused image regions are perceived, which is not the case for local texture perception. In the following section, we further analyze this interesting trend.

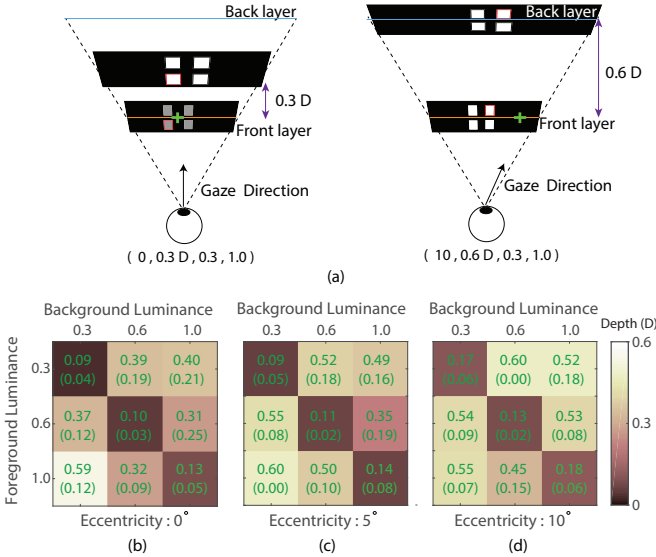



Fig. 5: Artifact perception at depth discontinuity. (a) The configuration of the perceptual experiment. The values in parentheses represents the eccentricity, depth difference, foreground luminance and background luminance. (b–d) The experimental results on the depth difference threshold at which LB and LFS are distinguishable. The mean values are shown with the standard deviations in parentheses.

### 6.2 Analysis of edge profiles

In order to clarify the occlusion perception, we investigate 1D luminance profiles (Fig. 6) that are produced at the fovea by LB and LFS

methods, while observing a depth edge between the front and back planes. We assume that the eye is always focused at the front plane, which leads to the strongest artifacts [37].

The E-1—E-3 types in Fig. 6 show the depth discontinuity where the luminance values are the same for the front and back planes. In such conditions the artifact patterns in the LB decomposition can be attributed to an interaction of two factors: optical blur in the back plane and luminance additivity in our two-plane display (Sect. 2.2). As the energy of the blurred signal increases with the back-plane luminance, the artifact absolute magnitude is larger in the E-3 than the E-1 case. However, the artifact detectability, akin to Weber's law, depends on its luminance contrast with respect to the uniform background; thus, the E-1—E-3 types have similar thresholds (Fig. 5 (b)). In general, the eye sensitivity for this type of artifact is relatively high, as the contrast detection thresholds at uniform background are relatively low [26]. The artifact contrast increases with depth discontinuity, so that it can easily be detected even for small depth differences (Fig. 5 (b)).

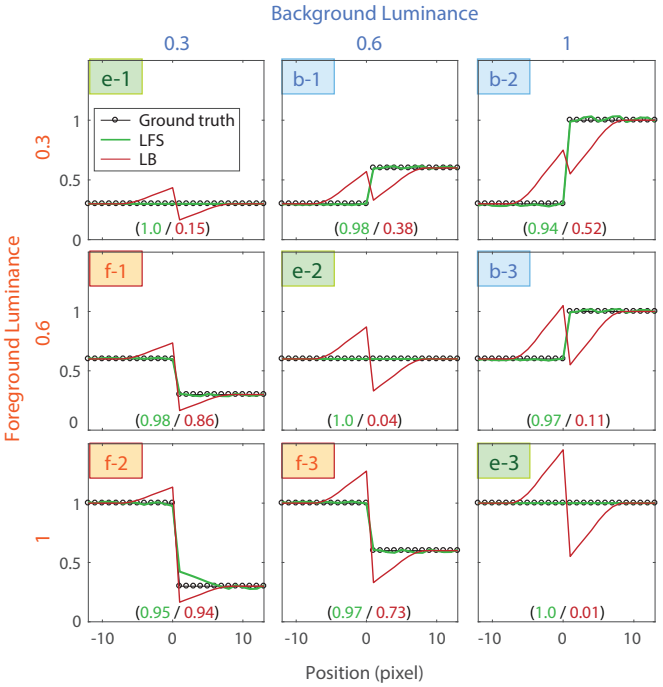



Fig. 6: 1D luminance profiles at the edge between the front and back planes separated with the 0.6 D depth difference for LB and LFS decompositions. The right half of each plot corresponds to the back plane, and the left half to the front plane, which is also the focus plane. At the bottom of each case, the SSIM value pairs of (LFS vs. ground truth / LB vs. ground truth) are shown in the parentheses. These SSIM values clearly indicate that LFS surpasses LB in all cases.

The F-1—F-3 types in Fig. 6 show the depth discontinuity where the front-plane luminance values are higher than their back counterpart. Similar artifact patterns as in the E-1—E-3 types are created, but this time they are imposed on contrast edges that act as contrast maskers [26]. Effectively contrast discrimination thresholds for such artifacts are elevated, which requires significant increase of depth discontinuity to make the artifact visible (Fig. 5 (b)).

The B-1—B-3 types in Fig. 6 show the depth discontinuity where the back-plane luminance values are higher than their front counterpart. This time the artifact pattern is embedded into the edge luminance profile, which might result in a more blurry edge appearance. Nevertheless, the HVS sensitivity for such artifact patterns is similar to the F-1—F-3 types (Fig. 5 (b)) with remarkably close depth thresholds for the same luminance contrast (the F-1 and B-1, and F-3 and B-3 types). This observation does not hold for the F-2 and B-2 types, and it can possibly be attributed to the imperfect luminance profiles for LFS due to intensity saturation caused in the constrained optimization. Interestingly, the variance in the participant responses is higher for the B-1—B-3 types than their F-1—F-3 counterparts.

While we do not conduct a detailed analysis for the eccentricity cases (Fig. 5 (c-d)), overall, similar observations can be made.

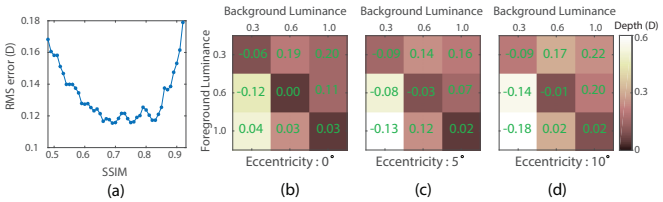



Fig. 7: SSIM calibration. (a) The RMS error between the predicted depth thresholds and experimental results. (b–d) the predicted depth difference thresholds from SSIM for various eccentricities. The values represent the error between the predicted thresholds and experimental outcome.

### 6.3 Calibrating SSIM

Similar to our previous analysis in Sect. 5.2, we calibrate the SSIM to predict the outcome of perceptual experiments (Sect. 6.1). Instead of using the previous detection threshold, we derive the optimal SSIM for detecting artifacts at occlusions independently. This strategy follows the observations made in [1,41], where specific training for each artifact type led to the improvement of SSIM metric predictions.

For each combination of the luminance contrast and depth difference, we generate the front focal images for both LFS and LB, and compute the minimum value in the SSIM map between the two algorithms. In order to simulate the perceived image in the peripheral vision, we apply the Gaussian blur with the cutoff frequency according to the quantitative HVS model by Watson [43]. The RMS error between the predicted and actual depth differences is the smallest, around 0.7–0.83 (Fig. 7(a)), and we conservatively select the largest value as the detection threshold. The depth difference thresholds as predicted by the SSIM are shown in Fig. 7 (b–d) for various eccentricities. The errors are typically acceptable when compared to the variance in the user experiment in Fig. 5 (b–d). However, the SSIM prediction produces depth thresholds that are consistently too large for the F-type distortions and too small for the B type (Fig. 6). This discrepancy in the SSIM sensitivity might be attributed to differences in the distortion profiles as discussed in Sect. 6.2. In further considerations, we rely on a more conservative prediction for the B type.

Using the calibrated SSIM, we can predict depth thresholds for larger eccentricities in display configurations of wider field of view and extended dioptric range. Based on these predictions, combined with the experiment outcome in the fovea and near eccentricity (Sect. 6.1), we investigate the selection rule for finding regions to apply LFS in the next section.

## 7 Unified Optimization

Our analyses reveal that LB can be applied for all textured surfaces (Sect. 5). For occlusion boundaries, we need to selectively apply LFS depending on the luminance contrast, depth difference and eccentricity (Sect. 6). In this section, we establish the selection rule for LFS based on the occlusion analysis and propose a unified optimization to integrate LB and LFS.

### 7.1 Selection rule

We design the selection rule for LFS as a function of the Michelson contrast, eccentricity and depth. We first express each combination of background and foreground luminance as Michelson contrast. In this case the F-1 and B-1, F-2 and B-2, F-3 and B-3, and E-1–E-3 types have the same contrast. Among two or three different depth thresholds for a given contrast, we select the smallest depth thresholds to be on the conservative side. In our SSIM prediction, we also aim to study the perception of artifacts at large eccentricities, which is expected to lead to larger depth thresholds. In order to check the depth separation

beyond 0.6 D, we simulate four-plane displays with a 0.6 D gap between successive layers. Our experimental outcome still holds for this display configuration since LB assigns the values to two nearby planes only; therefore, the behavior of LB in our display and the four-plane display is the same for edges with less than 0.6 D separation. In Fig. 8, we extrapolate the depth threshold to 50° eccentricity. Then, we fit a 3D surface to the predicted depth thresholds. The depth thresholds obtained from the perceptual experiments are marked with red points, and the predicted thresholds from SSIM are indicated with blue points. Although further confirmation is required with the perceptual experiments, our prediction suggests that a huge computational gain could possibly be obtained in wide field of view multi-layered displays in the future. In our implementation, we apply LFS to the cases where the depth difference is larger than the depth thresholds on the predicted surface.

Based on the selection rule, we generate a mask indicating regions that require LFS. The example of a mask generation for a fish scene in Fig. 12 is shown in Fig. 9. From the depth map (Fig. 9(a)) and Michelson contrast map (Fig. 9(b)), we generate a mask to apply LFS for the center gaze direction (Fig. 9(c)). The A and B cases show the occlusion boundaries eliminated from the mask due to the decreased sensitive at high eccentricities. The C case is an example of type E edges in Fig. 6. Although this edge has a small depth difference, it is still masked due to its lower luminance contrast compared to nearby edges.

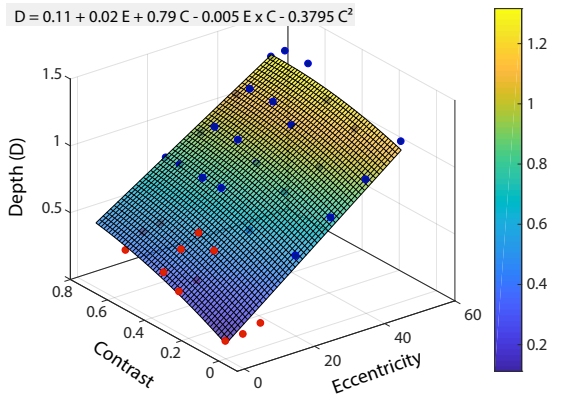
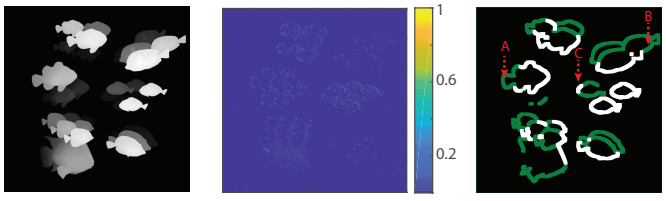



Fig. 8: The predicted depth difference thresholds from the SSIM. The goodness of fit: $R^2 = 0.9603$, RMSE= 0.068. For the measurement data only, $R^2 = 0.8078$, RMSE= 0.071.

### 7.2 Unified decomposition framework

We propose a unified optimization scheme which solves LFS with LB as a constraint. In practice, we can calculate LFS and LB separately and blend the results at intersection regions. However, keeping in mind that LFS requires a constrained least square optimization, using LB as the boundary condition for LFS can provide a smooth transition at intersections. The original decomposition algorithm of LFS can be written in the following form:




(a) (b) (c)

Fig. 9: Mask generation. (a) Depth map. (b) Michelson contrast map. (c) Mask. White region: masked region for the center gaze direction. Green region: masked region assuming no degradation of HVS at high eccentricities.

$$\underset{(KN)\times 1}{\begin{bmatrix}\mathbf{L}(v,u_1)\\ \mathbf{L}(v,u_2)\\ \vdots\\ \mathbf{L}(v,u_K)\end{bmatrix}} = \underset{(KN)\times(DN)}{\begin{bmatrix}\mathbf{P}_{11} & \mathbf{P}_{12} & \dots & \mathbf{P}_{1D}\\ \vdots & \vdots & \vdots & \vdots\\ \mathbf{P}_{K1} & \mathbf{P}_{K2} & \dots & \mathbf{P}_{KD}\end{bmatrix}} \underset{(DN)\times 1}{\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_D\end{bmatrix}}. \quad (1)$$

Here we employ a two-plane parametrization of the light field. $v$ denotes the spatial coordinate on the light field plane and $u$ denote the spatial position on the pupil. $\mathbf{L}(v,u_k)$ is a vectorized 2D image given a viewpoint $k$ and $\mathbf{x_d}$ is a vectorized 2D pixel value on the display layer $d$. $K$ is the number of viewpoints and $D$ is the number of layers. Without loss of generality, we assume that the number of pixels in each layer and target light field are both equal to $N$. In practice, the target light fields can have a different resolution. The submatrix $\mathbf{P}_{kd}$ of projection matrix $\mathbf{P}$ is defined as follows [25]: $(\mathbf{P}_{kd})_{i,j} = 1$ if $L(i,u_k)$ intersects with $(\mathbf{x}_d)_j$, and 0 otherwise.

We divide each component into two regions: the masked and unmasked regions. We apply full decomposition to the masked region, and the linear blending rule to the unmasked region. The subscript $M$ denotes "masked" and $U$ denotes "unmasked".

$$\mathbf{L}(v,u_k) = \begin{bmatrix}\mathbf{L}(v,u_k)_M\\ \mathbf{L}(v,u_k)_U\end{bmatrix}, P_{kd} = \begin{bmatrix}P_{kd,M}\\ P_{kd,U}\end{bmatrix}, \mathbf{x}_d = \begin{bmatrix}\mathbf{x}_{d,M}\\ \mathbf{x}_{d,U}\end{bmatrix} \quad (2)$$

Then the original equation can be rewritten as follows:

$$\underset{(KN)\times 1}{\begin{bmatrix}\mathbf{L}(v,u_1)_M\\ \mathbf{L}(v,u_1)_U\\ \mathbf{L}(v,u_2)_M\\ \mathbf{L}(v,u_2)_U\\ \vdots\\ \mathbf{L}(v,u_K)_M\\ \mathbf{L}(v,u_K)_U\end{bmatrix}} = \underset{(KN)\times(DN)}{\begin{bmatrix}P_{11,M} & P_{12,M} & \dots & P_{1D,M}\\ P_{11,U} & P_{12,U} & \dots & P_{1D,U}\\ P_{21,M} & P_{22,M} & \dots & P_{2D,M}\\ P_{21,U} & P_{22,U} & \dots & P_{2D,U}\\ \vdots & \vdots & \vdots & \vdots\\ P_{K1,M} & P_{K2,M} & \dots & P_{KD,M}\\ P_{K1,U} & P_{K2,U} & \dots & P_{KD,U}\end{bmatrix}} \underset{(DN)\times 1}{\begin{bmatrix}\mathbf{x_{1,M}}\\ \mathbf{x_{1,U}}\\ \mathbf{x_{2,M}}\\ \mathbf{x_{2,U}}\\ \vdots\\ \mathbf{x_{D,M}}\\ \mathbf{x_{D,U}}\end{bmatrix}} \quad (3)$$

Since the linear blending rule is applied for a single image, we eliminate $\mathbf{L}(v,u_k)_U, P_{kd,U}$ and $\mathbf{x}_{d,U}$ for $k > 1$, assuming that $u_1$ is the center viewpoint. Then, the unmasked region is handled only with the center viewpoint, $\mathbf{L}(v,u_1)_U$.

$$\underset{(N+(K-1)N_M)\times 1}{\begin{bmatrix}\mathbf{L}(v,u_1)_M\\ \mathbf{L}(v,u_1)_U\\ \mathbf{L}(v,u_2)_M\\ \vdots\\ \mathbf{L}(v,u_K)_M\end{bmatrix}} = \underset{(N+(K-1)N_M)\times(DN)}{\begin{bmatrix}P_{11,M} & P_{12,M} & \dots & P_{1D,M}\\ P_{11,U} & P_{12,U} & \dots & P_{1D,U}\\ P_{21,M} & P_{22,M} & \dots & P_{2D,M}\\ \vdots & \vdots & \vdots & \vdots\\ P_{K1,M} & P_{K2,M} & \dots & P_{KD,M}\end{bmatrix}} \underset{(DN)\times 1}{\begin{bmatrix}\mathbf{x_{1,M}}\\ \mathbf{x_{1,U}}\\ \mathbf{x_{2,M}}\\ \mathbf{x_{2,U}}\\ \vdots\\ \mathbf{x_{D,M}}\\ \mathbf{x_{D,U}}\end{bmatrix}} \quad (4)$$

By applying the linear blending, we can reduce the dimension of the projection matrix from $(KN)\times(DN)$ to $(N+(K-1)N_M)\times(DN)$, where $N_M$ denotes the number pixels in the masked region. For example, if $K=9, D=3, N_M = N/4$, then the dimension changes from $(9N)\times(3N)$ to $(3N)\times(3N)$.

Solving the reduced decomposition problem, however, does not provide the correct answer because $\mathbf{x}_{d,U}$ do not have enough constraints. The multi-viewpoint images impose constrains on each pixel value, but the single viewpoint cannot. Therefore, the pixel values should be calculated separately according to the linear blending rule and replaced in each iteration step.

## 8 Implementation

### 8.1 Rendering and decomposition

Our rendering pipeline breaks down to four steps: (1) rendering the central viewpoint image and depth map, (2) computing the mask for LFS, (3) rendering additional viewpoint images on the masked region,

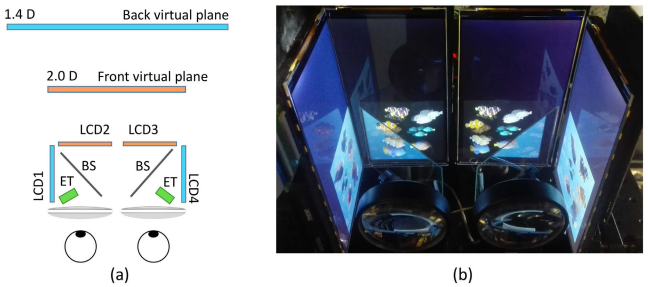



Fig. 10: The display prototype. (a) The schematic and (b) photograph of the display system. BS:Beam splitter, ET:Eye tracker.

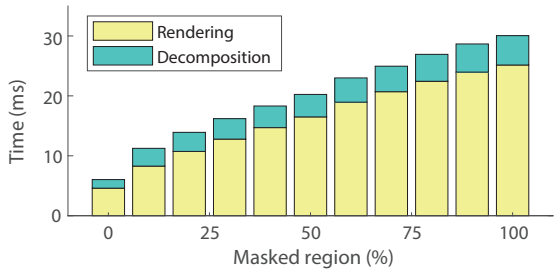



Fig. 11: The rendering and decomposition timings for various ratios of masked region for the fish scene in Fig. 12.

and (4) performing LB and the iterative decomposition using SART. In the case of full light field synthesis without any mask, we render 9 views with a $1200 \times 1200$ resolution with a 1 ray/pixel. For our optimal decomposition, we first render a single 2D image and depth map. By analyzing the luminance contrast and depth gradient map, we calculate the masked region based on the criteria in Sect. 7.1. Next, the 8 viewpoint images except the center images falling inside the pupil are generated only for the masked region. Compared to the generation of full light fields, our selective rendering can greatly reduce the computation time. From the rendered target scene, we calculate the optimal decomposed images using the unified decomposition framework (Sect. 7.2). At this stage, we also reduce the optimization time over the conventional SART implementation by developing an efficient adaption of SART in CUDA. More details can be found in Supplementary Information C. The rendering system is implemented using Nvidia OptiX ray tracer, which enables the selective rendering for a given mask with minimal overhead. The renderer is driven by a PC with a 3.60 GHz Xeon CPU and 32.0 GB RAM equipped with a single Nvidia GTX 1080 TI graphics card.

### 8.2 Eye-tracked multi-layered accommodative display

We build a two-plane VR display to test the rendering strategy. The schematic and photograph of the setup are shown in Fig. 10. For each eye, images from two $2560 \times 1440$ LCD displays (Topfoison TF60010A) are combined with a beam splitter (Edmund Optics #64-408) and magnified with an achromatic lens (Thorlabs AC508-080-A). Eye-trackers (Pupil Labs) are placed right behind the two lenses. The optical system for the right eye is mounted on the linear stage for adjusting the interpupillar distance. The dioptric distances to the front and back virtual planes are set to 2.0 D and 1.4 D, respectively.

The resolution of display is $1200 \times 1200$, which is significantly higher than the light field displays reported so far [16, 24, 35]. FOV is $40^\circ$, and the angular resolution of the system is 15 cpd. Our system has a high enough resolution and large enough FoV to study the effect of foveation, while the resolution of current VR and AR systems rarely exceeds 10 cpd, which is quite limited for foveated rendering.

## 9 Results

We render three different scenes to test our rendering strategy. We first evaluate the computational time for our optimization algorithms. Then, we compare the visual quality of our method with LB and LFS on our display prototype and using simulations.

| Scene | # polygons | mask(%) | rendering | decomposition |
|---|---|---|---|---|
| Fish | 20498 | 7.3 | 9.26 (27.48) | 2.57 (4.11) |
| Dice | 569810 | 6.5 | 14.11 (47.08) | 2.44 (4.12) |
| Forest | 16924 | 1.8 | 7.29 (28.31) | 2.35 (4.19) |

Table 1: The rendering and decomposition timings of our hybrid method for various scenes. The rendering and decomposition timings are given in ms. The values in the parentheses indicate timings for full LFS rendering.

### 9.1 Performance

We measure the total rendering time for the whole pipeline during monocular viewing. For three scenes in Fig. 12, the rendering and decomposition timings for our hybrid method and full LFS are measured in Table 1. All decompositions are performed with 10 iterations. As the shader/geometry complexity becomes higher, the rendering time increases. However, the computational saving of our hybrid method is even more pronounced with respect to full LFS, since in many scene regions we require only a single view for LB and can avoid full LF rendering. We also observe that the decomposition time only depends on the percentage of masked region. Our test scenes contain 5.19% of LFS region on average. The frame rates are measured as 84 Hz ($\times$4.25), 60 Hz ($\times$4.06), 103 Hz ($\times$4.50) for the fish, dice and forest scenes. The values in the parentheses denote the speed enhancement over full LFS after subtracting a fixed cost of a single view rendering. If a scene contains many depth edges, the performance gain of our hybrid method reduces since most of the regions should be rendered with LFS. For binocular viewing conditions, the stereoscopic scenes are rendered sequentially. In this case, the total rendering time increases by a factor of 2.

In order to test the effect of percentage of masked region, we also measured timing for various masked region for the fish scene as shown in Fig. 11. Here, we generated the masks randomly, instead of using the mask generated by the selection rule in Sect. 7.1. The zero percentage corresponds to the LB-only rendering. The total optimization time linearly increases as the masked region grows. This trend implies that there is minimal overhead coming from selective rendering for randomly masked region.

### 9.2 Comparison

For each scene, we capture the photographs from our display prototype and simulate the perceived images as shown in Fig. 12. We compare three algorithms: LB, full LFS and our optimization. The masks are computed assuming the gaze is directed towards the center. For the simulated images, we compute the SSIM against ground truth, which is the focal image generated with dense light fields. The SSIM maps indicate that LB produces strong artifacts mostly along the occlusion boundaries. Furthermore, the boundaries between LFS and LB in our algorithm do not produce any noticeable discontinuities, confirming the validity of the unified decomposition framework. However, halo effects are visible around edges in captured images for both LFS and our method. We found that small errors in color calibration between two display layers led to such artifacts, which are not visible in the simulation.

The fish and dice scenes show various aspects of edge reconstruction analyzed in Fig. 6. The blue fish and gray dice are examples of the E-3- and E-2-type occlusions. LB generates a sharp contrast while LFS produces smooth transitions. On the other hand, as seen around the yellow fish and reddish-brown dice, the B-2-type edges from LB look blurry, but sharp edges are obtained in our method. The edges along the white part of the fish or the orange dice present F-2-type profiles. Since the foreground objects are brighter, the differences among the three algorithms are less obvious.

The forest scene demonstrates the reconstruction quality in textured regions. The high-frequency features of slanted grass fields are preserved in LB and our method, but they are blurred out in LFS, which is expected from Fig. 4(e). Although our method provides better image quality, this enhanced contrast could possibly lead to incorrect contrast

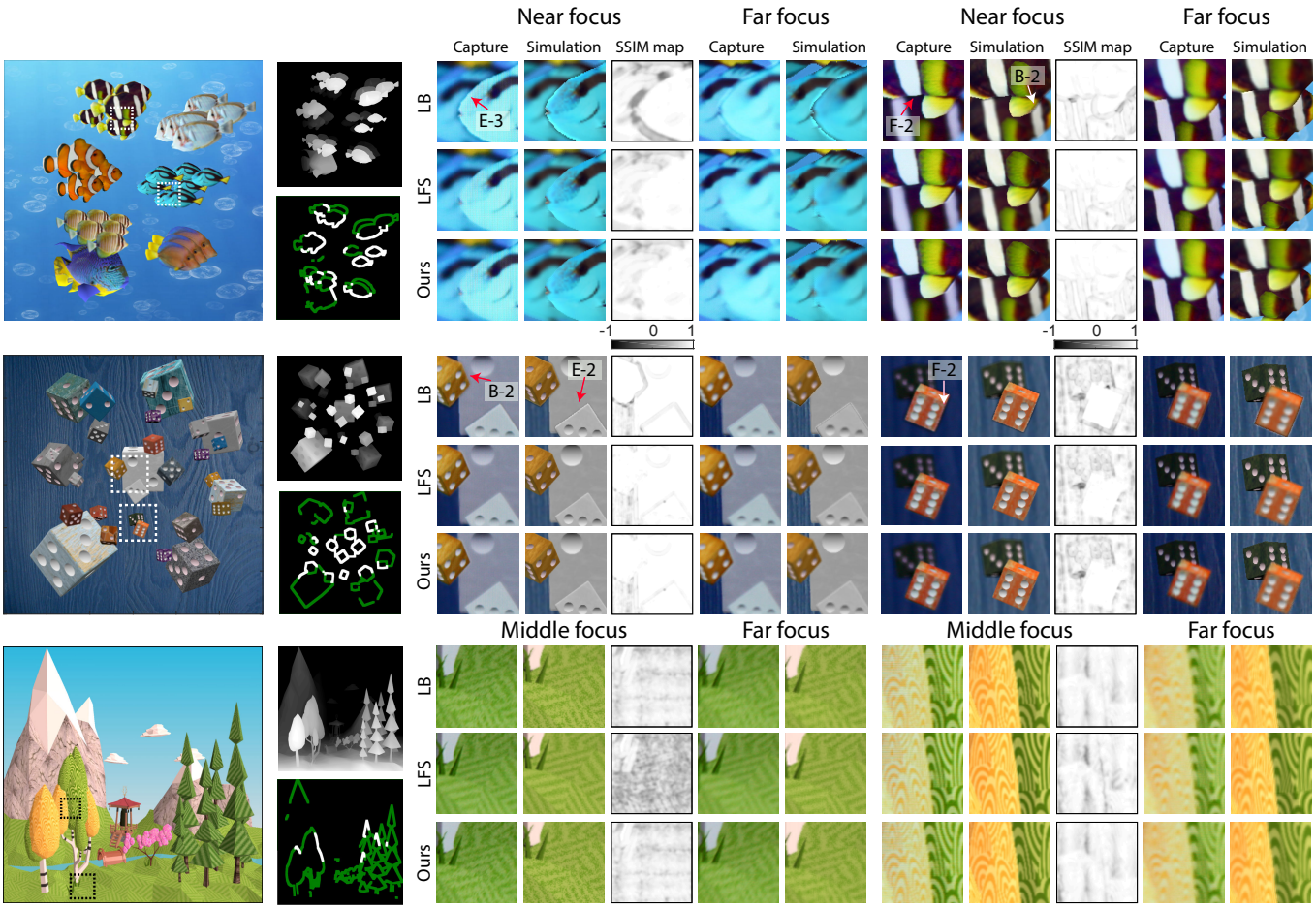



Fig. 12: Comparison of various decomposition methods for various scenes. The images (column 1) represent the target scenes. The upper and bottom images (column 2) are the corresponding depth maps and masks (white region: masked region for the center gaze direction, green region: masked region assuming no degradation of HVS at high eccentricities). For each scene, we compare ours (row 3) to the LB (row 1) and full LFS (row 2). The near or middle focus images (columns 3–5, 8–10) show captured images from the display, simulated perceived image, and SSIM map between the ground truth and simulation. The captured and simulated images at far focus are also shown in columns 6–7 and 11–12.

curves as seen in Fig. 2. However, far-focus images still look more blurry than the focused images on grass field, which suggests that the failure of LB at high spatial frequencies does not affect the effectiveness for driving accommodation [29]. On the other hand, low frequency textures on the yellow and green trees are reconstructed with slightly higher contrast for LFS, which is expected from the low frequency region in Fig. 4(d). However, for those regions the differences between the two methods are subtle, so they are not distinguishable according to our perceptual experiment and SSIM predictions.

### 9.3 Temporal coherence

We test the temporal coherence of our method on dynamic scenes, which is a critical use case for real-time methods. As our perceptual findings allowed us to consistently use LB for textured regions, temporal changes occur only around edges. Since the threshold functions on edges are derived based on indistinguishability between LB and LFS (Sect. 7.1), smooth transitions can be achieved when a switch between LFS and LB occurs near the threshold. We first test the transition behaviors in two dynamic scenarios. The results during object and camera motion are given as Scenes 1 and 2 in Supplementary Video, respectively. In both cases, we assume a gaze direction towards the center, which is marked with a red box. The captured and simulated videos do not show any noticeable artifacts around the edges near the gaze position. In the periphery, the transition between the two algorithms are sometimes visible, but those boundaries are not noticeable in actual viewing conditions due to the reduced sensitivity of HVS. In Scene 1, we observe rendering artifacts around the high spatial frequency textures which originate from the low sampling rate used (1 ray per pixel) and are unrelated to our mask quality. In order to address this issue, space-time ray-tracing methods can be employed in the future for a better rendering quality [9]. For a better visualization of various artifacts, we also computed the SSIM map between the images generated with our method and the ground truth. Here, the ground truth is computed as focal images generated with dense light fields. The SSIM maps indicate that high spatial frequency textures show noticeable deviations due to the use of single ray per pixel and imperfect reconstruction of high spatial frequencies as discussed in Fig. 2. The SSIM videos also show an error at occlusion boundaries in the periphery, but it is not noticeable due to the foveation. Although those artifacts are clearly seen in the SSIM maps, the rendered videos do not exhibit significant artifacts when they are observed alone without the comparison against the ground truth.

We additionally test mask stability during the use of an actual eye tracker (Scene 3) in Supplementary Video in order to see how our method behaves in the presence of measurement noise in gaze position. Due to physical constraints, it is not possible to capture a scene while an observer is using the display. Therefore, we captured the results using gaze positions recorded from an actual viewing session of an observer. In order to improve the stability of our method, we applied a simplified version of a denoising method proposed by Kumar [20] for gaze inputs. No visible artifacts were detected during our visual inspection, which indicates that our hybrid method is able to perform well when an eye tracker is used.

## 10 Evaluation

In order to validate the perceptual quality of our method, we conducted a user experiment to compare (LB, Ours) and (gaze-contingent LFS, Ours) for four static scenes in a binocular setting. In order to simulate the gaze-contingent sampling in Sect. 4, we show gaze direction stimuli and project decomposed images optimized for a given gaze direction. To allow the accommodation change, the gaze direction stimuli were set to rectangular boxes extending $2^\circ$ of visual angles for both eyes. This small gaze change does not introduce the generation of new decomposed images in gaze-contingent LFS. The subjects are instructed to maintain the gaze direction inside the box, but to judge the overall image quality. In each trial, the users are asked to choose the scene which produces better image quality. In each scene, the users compare the quality of scenes at five different gaze directions. Six subjects participated in the experiment. The outcome of perceptual evaluation is shown in Fig. 13. We observe high preference for our method over the LB and the difference between the two methods is found statistically significant in binomial test ($p < 0.05$ for all scenes). This can be attributed to the better reconstruction of edges in our method. In the forest scene, the difference between LB and our algorithm decreases since the scene mostly consists of textured regions without occlusion. In the comparison with LFS, our algorithm shows similar preferences for fish and dice scenes ($p > 0.50$), which indicate the observers are indifferent between the two methods. Considering the fact that those two scenes contain many occlusion boundaries, the outcome of user study suggest that our masking algorithm in Sect. 7.1 successfully works. However, the subjects preferred LFS over our method in the forest scene which contains high frequency features and the difference is significant ($p < 0.05$). Although we did not survey formally, subjects reported that they prefer blurred texture in LFS over ours (Fig. 12). Since the reconstruction of high frequency textures requires precise alignment, small pupil and head movements may lead to perception of those features as noise patterns.

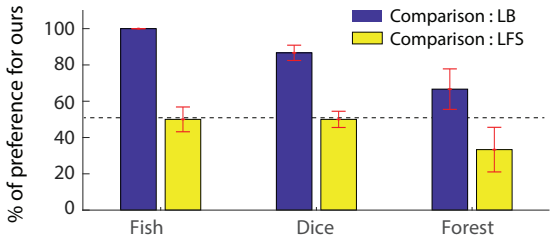



Fig. 13: The percentage of answers preferring our method over LB and LFS for three scenes. The error bars indicate the standard error.

## 11 Limitations

**Display** Our optical system is based on magnifier lenses; therefore, the system suffers from aberrations around the outer regions. The multi-plane displays with holographic optical elements [24] could be a good candidate for reducing the image degradation. Although our design provides a relatively wide FOV of $40^\circ$, it is still smaller than current VR displays. The dioptric range of the display is also limited in two-plane displays. Therefore, further validating our occlusion analysis in wide FoV multi-plane displays with larger dioptric range is required. The development of the displays with extended dioptric range also enables the study of foveation rules on the occlusion boundaries in the context of defocus states, while only eccentricity has been considered in our study. Similar to blurred artifacts at high eccentricities, large depth differences between the eye focus and edge can also reduce sensitivity to edge artifacts. Our prototype currently lacks devices measuring the accommodation states [18, 35]. The evaluation of the effectiveness of driving accommodation with different optimizations would be important future work.

**Perception** In this work, we relied on a specific image quality metric (SSIM). As the image quality evaluation is still an open problem, our detailed masking aggregation could be affected by SSIM inaccuracies. As observed in the validation experiment, the prediction of SSIM does not account for artifacts induced by viewing conditions such as pupil movements and misalignment. Since the perceived images depend greatly on the focal state, a quality metric that can meaningfully compare lightfields would be required. Such a metric should be applied after considering display specific limitations in reproducing light fields. Also, a metric capable of predicting the ability to induce the eye accommodation by such reproduced light fields would be desirable in deriving possibly new foveation rules in our approach. We relegate all these interesting and difficult problems to future work.

**Rendering** In this work, we focus on Lambertian scenes, and handling glossy objects would require the extension of our masking algorithm to consider such objects as a function of the visibility of view-dependent effects. Since the boundaries between LFS and LB show smooth transitions as seen in Fig. 12, extending the mask region should handle non-Lambertian scenes as well. We also did not perform a direct a comparison to RO method. Although RO performs moderately better at 9 and 12 cpd according to our analysis in Fig. 2, this quality improvement comes at the significant computational speed loss. It is noteworthy that the rendering speed of RO is reported as 5 FPS for a display resolution of $512\times512$ [35], while the performance time of our algorithm is faster than 60 FPS for a display resolution of $1200\times1200$. Furthermore, none of the methods can correctly trigger accommodation at 12 cpd, therefore we concluded that the gaze-contingent LFS is a suitable method providing similar quality offered by RO yet with much faster computational speed. Although we expect performance gain without significant quality degradation, we leave this comparison for future work. We also believe that our strategy can be successfully used to combine RO and LB techniques, but this requires further investigation.

## 12 Conclusions

In this paper, we developed a hybrid decomposition framework of the linear blending and the light field synthesis enabling the real-time rendering and high-fidelity reconstruction in multi-layered light field displays. Our perceptual experiments and the SSIM analysis provide a deeper insight into visual quality produced by different decomposition algorithms. In particular, we show that for textured surfaces, LB and LFS are indistinguishable for low to mid spatial frequencies, and LB is closer to the ground truth for high spatial frequencies. For occlusion boundaries, LB fails at low luminance contrast edges rather than high contrast edges, which seems counterintuitive but is a consequence of additive combining of focused and defocused patterns at both edge sides. We also show that those conditions for occlusions can be further relaxed for surfaces at sufficiently large eccentricities, when the sensitivity of the HVS drops significantly. In order to apply our selective optimization strategy, we develop a unified optimization framework of LB and LFS. We tested our optimal rendering strategy with a two-layer multi-plane display and validate the 60 Hz rendering time for $1200\times1200$ resolution with 9 viewpoints.

While our algorithm focuses on the additive light-field display, our hybrid strategy can possibly be extended to the multiplicative light-field displays since it can be formulated with the additive light-field synthesis under logarithm [23]. Therefore, investigating the simple decomposition rules in a multiplicative architecture and integrating with the light field synthesis algorithms would be an interesting topic. Since the major artifacts of LB around the occlusion originate from the additive nature of our displays, studying the edge artifact in the multiplicative display would lead to interesting perceptual insights in accommodative light-field displays.

The perceptual evaluation of optimization algorithms for dynamic scenes could be another interesting topic of future work. Even though the incorrect boundaries of the linear blending are clearly visible in static scenes, it is unclear whether artifacts are noticeable under motion blur. Therefore, studying the perception of artifacts in interactive and dynamic scenes could provide additional computational benefits.

## Acknowledgments

The authors would like to thank Oliver Mercier for helpful discussions. The project was supported by the Fraunhofer and Max Planck

cooperation program within the German pact for research and innovation (PFI). This project has also received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie grant agreements No 642841 (DISTRO) and from the European Research Council (ERC) (grant agreement No 804226/PERDY). The project was partially funded by the Polish National Science Centre (decision number DEC-2013/09/B/ST6/02270).

## REFERENCES

[1] V. K. Adhikarla, M. Vinkler, D. Sumin, R. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk. Towards a quality metric for dense light fields. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[2] K. Akeley, S. J. Watt, A. R. Girshick, and M. S. Banks. A stereo display prototype with multiple focal distances. *ACM Transactions on Graphics*, 23(3):804, aug 2004.

[3] K. Aksit, 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, nov 2017.

[4] A. Andersen. Simultaneous Algebraic Reconstruction Technique (SART): A superior implementation of the ART algorithm. *Ultrasonic Imaging*, 6(1):81–94, jan 1984.

[5] M. S. Banks, D. M. Hoffman, J. Kim, and G. Wetzstein. 3D Displays. *Annual Review of Vision Science*, 2(1):397–435, 2016.

[6] S. A. Cholewiak, G. D. Love, P. P. Srinivasan, R. Ng, and M. S. Banks. Chromablur: Rendering chromatic eye aberration improves accommodation and realism. *ACM Trans. Graph.*, 36(6):210:1–210:12, Nov. 2017.

[7] T. F. Coleman and Y. Li. A Reflective Newton Method for Minimizing a Quadratic Function Subject to Bounds on Some of the Variables. *SIAM Journal on Optimization*, 6(4):1040–1058, nov 1996.

[8] D. Dunn, C. Tippets, K. Torell, P. Kellnhofer, K. Aksit, 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*, 23(4):1322–1331, apr 2017.

[9] A. S. Glassner. Spacetime ray tracing for animation. *IEEE Computer Graphics and Applications*, (2):60–61, 1988.

[10] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3D graphics. *ACM Transactions on Graphics*, 31(6):1, nov 2012.

[11] F. Heide, G. Wetzstein, R. Raskar, and W. Heidrich. Adaptive image synthesis for compressive displays. *ACM Transactions on Graphics*, 32(4):1, jul 2013.

[12] X. Hu and H. Hua. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. *Optics express*, 22(11):13896–903, 2014.

[13] H. Hua. Enabling Focus Cues in Head-Mounted Displays. *Proceedings of the IEEE*, 105(5):805–824, may 2017.

[14] H. Hua and B. Javidi. A 3D integral imaging optical see-through head-mounted display. *Optics Express*, 22(11):13484, jun 2014.

[15] F.-C. Huang, K. Chen, and G. Wetzstein. The light field stereoscope. *ACM Transactions on Graphics*, 34(4):60:1–60:12, jul 2015.

[16] F.-C. Huang, D. Luebke, and G. Wetzstein. The light field stereoscope. In *ACM SIGGRAPH 2015 Emerging Technologies on - SIGGRAPH '15*, pages 1–1, New York, New York, USA, 2015. ACM Press.

[17] R. Konrad, N. Padmanaban, K. Molner, E. A. Cooper, and G. Wetzstein. Accommodation-invariant computational near-eye displays. *ACM Transactions on Graphics*, 36(4):1–12, jul 2017.

[18] G.-A. Koulieris, B. Bui, M. S. Banks, and G. Drettakis. Accommodation and Comfort in Head-Mounted Displays. *ACM Transactions on Graphics*, 36(4):1–11, 2017.

[19] G. Kramida. Resolving the Vergence-Accommodation Conflict in Head-Mounted Displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, jul 2016.

[20] M. Kumar, J. Klingner, R. Puranik, T. Winograd, and A. Paepcke. Improving the accuracy of gaze input for interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ETRA '08, pages 65–68, New York, NY, USA, 2008. ACM.

[21] M. Lambooij, W. IJsselsteijn, M. Fortuin, and I. Heynderickx. Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. *Journal of Imaging Science and Technology*, 53(3):030201, 2009.

[22] D. Lanman and D. Luebke. Near-eye light field displays. *ACM Transactions on Graphics*, 32(6):1–10, nov 2013.

[23] D. Lanman, G. Wetzstein, M. Hirsch, W. Heidrich, and R. Raskar. Polarization fields. *ACM Transactions on Graphics*, 30(6):1, dec 2011.

[24] 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 (Invited Paper). *IEEE Access*, 4(c):1–1, 2017.

[25] S. Lee, C. Jang, S. Moon, J. Cho, and B. Lee. Additive light field displays. *ACM Transactions on Graphics*, 35(4):1–13, jul 2016.

[26] G. Legge and J. Foley. Contrast masking in human vision. *Journal of the Optical Society of America*, 70(12):1458–1471, 1980.

[27] A. Levin and F. Durand. Linear view synthesis using a dimensionality gap light field prior. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1831–1838. IEEE, jun 2010.

[28] G. D. Love, D. M. Hoffman, P. J. W. Hands, J. Gao, A. K. Kirby, and M. S. Banks. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Optics express*, 17(18):15716–25, 2009.

[29] K. J. MacKenzie, D. M. Hoffman, and S. J. Watt. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. *Journal of Vision*, 10(8):22–22, jul 2010.

[30] A. Maimone and H. Fuchs. Computational augmented reality eyeglasses. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, number October, pages 29–38. IEEE, oct 2013.

[31] A. Maimone, A. Georgiou, and J. S. Kollin. Holographic Near-Eye Displays for Virtual and Augmented Reality. *ACM Transactions on Graphics*, 36(4):1–16, 2017.

[32] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, July 2011.

[33] S. Mathews and P. B. Kruger. Spatiotemporal transfer function of human accommodation. *Vision Research*, 34(15):1965–1980, aug 1994.

[34] N. Matsuda, A. Fix, and D. Lanman. Focal surface displays. *ACM Transactions on Graphics*, 36(4):1–14, 2017.

[35] 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):1–15, 2017.

[36] S. Moon, C.-K. Lee, D. Lee, C. Jang, and B. Lee. Layered Display with Accommodation Cue using Scattering Polarizers. *IEEE Journal of Selected Topics in Signal Processing*, 4553(c):1–1, 2017.

[37] 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):59:1–59:12, 2015.

[38] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

[39] S. Ravikumar, K. Akeley, and M. S. Banks. Creating effective focus cues in multi-plane 3D displays. *Optics Express*, 19(21):20940, 2011.

[40] Q. Sun, F.-C. Huang, J. Kim, L.-Y. Wei, D. Luebke, and A. Kaufman. Perceptually-guided foveation for light field displays. *ACM Transactions on Graphics*, 36(6):1–13, nov 2017.

[41] N. T. Swafford, J. A. Iglesias-Guitian, C. Koniaris, B. Moon, D. Cosker, and K. Mitchell. User, metric, and computational evaluation of foveated rendering methods. In *Proceedings of the ACM Symposium on Applied Perception*, SAP '16, pages 7–14, 2016.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[43] A. B. Watson and A. J. Ahumada. Blur clarified: A review and synthesis of blur discrimination. *Journal of Vision*, 11(5):10, 2011.

[44] A. B. Watson and D. G. Pelli. Quest: A bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2):113–120, Mar 1983.

[45] G. Wetzstein, D. Lanman, W. Heidrich, and R. Raskar. Layered 3D. *ACM Transactions on Graphics*, 30(4):1, jul 2011.

[46] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. K. Mantiuk. Dataset and metrics for predicting local visible differences. *ACM Transactions on Graphics (TOG)*, 37(5):172, 2018.

[47] H.-J. Yeom, H.-J. Kim, S.-B. Kim, H. Zhang, B. Li, Y.-M. Ji, S.-H. Kim, and J.-H. Park. 3D holographic head mounted display using holographic optical elements with astigmatism aberration compensation. *Optics Express*, 23(25):32025, 2015.

[48] M. Zannoli, G. D. Love, R. Narain, and M. S. Banks. Blur and the perception of depth at occlusions. *Journal of Vision*, 16(6):17, apr 2016.