Optimal Presentation of Imagery with Focus Cues on Multi-Plane Displays

Rahul Narain[1] Rachel A. Albert[2] Abdullah Bulbul[2] Gregory J. Ward[3] Martin S. Banks[2] James F. O'Brien[2]

[1]University of Minnesota [2]University of California, Berkeley [3]Dolby Laboratories

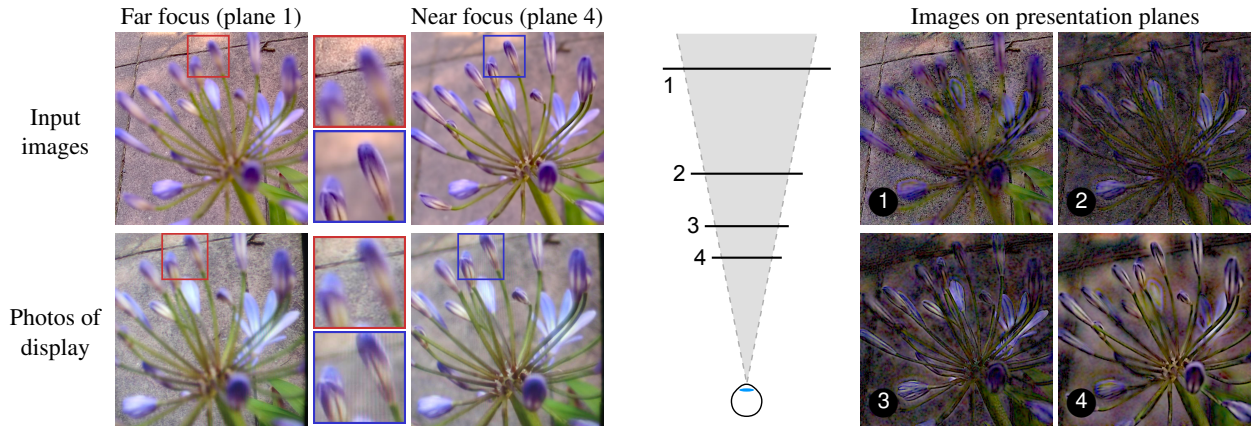


Figure 1: *Reproducing a real-world scene on a multi-plane display. Given a focus stack consisting of images of a scene focused at different distances, we use optimization to determine images to show on the presentation planes of the multi-plane display so that the image seen through the display when focusing at different distances matches the corresponding image of the input scene. The presentation planes combine additively in the viewer's eye to produce an image with realistic focus cues.*

Abstract

We present a technique for displaying three-dimensional imagery of general scenes with nearly correct focus cues on multi-plane displays. These displays present an additive combination of images at a discrete set of optical distances, allowing the viewer to focus at different distances in the simulated scene. Our proposed technique extends the capabilities of multi-plane displays to general scenes with occlusions and non-Lambertian effects by using a model of defocus in the eye of the viewer. Requiring no explicit knowledge of the scene geometry, our technique uses an optimization algorithm to compute the images to be displayed on the presentation planes so that the retinal images when accommodating to different distances match the corresponding retinal images of the input scene as closely as possible. We demonstrate the utility of the technique using imagery acquired from both synthetic and real-world scenes, and analyze the system's characteristics including bounds on achievable resolution.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—[Display Algorithms]

Keywords: Computational displays, multi-plane displays, eye accommodation, retinal blur, vergence-accommodation conflict

From the conference proceedings of ACM SIGGRAPH 2015.
Appearing in ACM Transaction on Graphics Vol. 34, No. 4.

1 Introduction

The human visual system uses a number of different cues to estimate the third dimension from the 2D retinal images. Some of these—e.g. shading, perspective, and occlusion—can be reproduced in a single 2D image shown on a conventional monitor. However, other important depth cues cannot be shown on conventional displays because they arise from the geometrical relationship between scene objects and the optics of the eyes. When observing an object in a natural scene, the viewer must adjust the angle between the two eyes' lines of sight (*vergence*) to fuse the object's images in the two eyes; simultaneously, each eye adjusts the focal power of its lens (*accommodation*) to create a sharp retinal image of the object. In natural scenes, these cues are consistent, because the vergence distance, where the lines of sight intersect, and the accommodation distance, where objects are in focus, are both equal to the optical distance of the fixated object.

To display a 3D scene with vergence cues requires presenting different scenes to each eye to produce binocular disparity, while correct accommodation cues require that blur corresponds to focus changes as viewers accommodate to different distances in real time. Stereoscopic displays that provide vergence cues are now commonplace in movie theaters and are commercially available in consumer televisions. However, despite the varying depth indicated by the vergence cues in these displays, the accommodation distance to produce a sharp image remains fixed at the distance to the display surface. This *vergence-accommodation conflict* results in perceptual distortions [Watt et al. 2005], difficulty in simultaneously fusing and focusing the image [Akeley et al. 2004; Hoffman et al. 2008], and viewer discomfort and fatigue in long-term use [Emoto et al. 2005; Hoffman et al. 2008; Lambooij et al. 2009; Shibata et al. 2011]. For the displays of the future to allow effective, comfortable, and realistic viewing of stereoscopic images, they must also support correct accommodation and defocus effects.

Light field displays and volumetric displays have recently been de-

veloped with the capacity to present correct accommodation cues. Light field displays modulate emitted light as a function of both position and direction, reproducing vergence cues as well as parallax due to head motion. With increasing angular resolution, such displays will also be able to produce correct cues to accommodation [Maimone et al. 2013], although at present this capability has not been demonstrated for human viewers. Volumetric displays, on the other hand, present three-dimensional imagery by placing light sources at multiple optical distances, so that accommodative effects arise automatically. Stereoscopic multi-plane display architectures [Akeley et al. 2004; Love et al. 2009; Liu et al. 2010], in particular, can present high-resolution imagery on a discrete set of *presentation planes* at different accommodative distances. Previous studies have shown that these displays can provide the appropriate stimuli to drive accommodation for simple, isolated, diffuse objects at intermediate distances [MacKenzie et al. 2010; Ravikumar et al. 2011]. However, it has not been clear how to present arbitrary 3D scenes—i.e., those including occlusions and reflections—with accurate accommodation cues on the additive layers of multi-plane displays.

We show how to improve the reproduction of accommodation cues on current hardware through computational methods. Rather than optimizing for a theoretical pinhole viewer by resampling scene data to neighboring display elements, we argue that one should distribute light in a way that takes into account the defocus that will occur in a human viewer's eye. Our specific contribution is a novel computational technique to optimally present arbitrary scenes on a multi-plane display, accurately reproducing the defocus behavior of occlusions, reflections, and other nonlocal effects as a function of accommodation. We take as input a focal stack of images representing the desired views of the scene when focusing at different distances, and use optimization to determine the images to display on the presentation planes so that the view through the display best matches the input for all accommodation distances. This approach makes it possible to deliver nearly correct accommodation cues in general scenes with sufficient accuracy for comfortable viewing.

2 Related work

There have been many recent developments in 3D display technology. A detailed discussion can be found in the reviews by Wetzstein et al. a[2012a] and Masia et al. [2013]. Below we give a brief overview, emphasizing accommodative effects.

Stereoscopic displays typically use a single display screen along with specialized glasses to present different images to the viewer's left and right eyes. A survey of hardware techniques for such displays is given in the SIGGRAPH course by Hirsch and Lanman [2010]. These displays provide appropriate vergence cues for depth, but the accommodative distance remains fixed at the distance to the screen. Thus appropriate correlation between vergence and accommodation, and between their sensory analogs of disparity and blur, is generally not possible with these displays. Fatigue effects due to vergence-accommodation conflict are well known; content authors often try to minimize them by composing scenes so that the main subject of the scene is presented with zero disparity [Mendiburu 2009]. Research motivated by this problem includes the work of Lang et al. [2010] who develop a computational approach for disparity mapping, and Du et al. [2013] who present a statistical model of discomfort based on disparity, motion, and spatial frequency. The zero-disparity heuristic limits scene composition and still produces conflict when other objects in the scene are fixated. Additionally, defocus effects due to finite aperture must be statically included in the presented images. As a result, when viewers fixate on blurred objects they may be able to fuse the stereo images and accommodate to the display, but the objects will remain blurred. This situation creates artifacts and incorrect scale cues [Held et al. 2010].

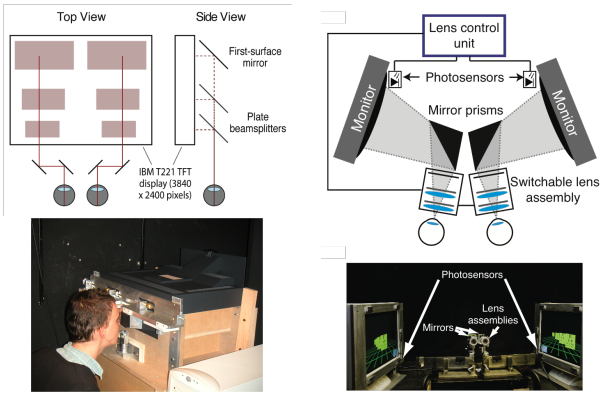
Light field displays are designed to reproduce a given four-dimensional light field, allowing unencumbered glasses-free viewing with both vergence and parallax. Initial approaches used lenticular arrays [Lippmann 1908; Matusik and Pfister 2004] and parallax barriers [Ives 1903; Perlin et al. 2000] to direct exitant light along different rays. Later developments have explored compressive techniques based on multi-layer architectures [Lanman et al. 2010; Wetzstein et al. 2011; Lanman et al. 2011; Wetzstein et al. 2012b]. Huang et al. [2012; 2014] have used similar techniques to design displays that correct for defocus and other aberrations in viewers with imperfect vision.

In principle, an ideal infinite-resolution autostereoscopic display could also produce accurate accommodation cues, because a light field theoretically encodes the full radiance distribution emitted from the scene. However, for normal viewing distances, presenting accommodation cues to human viewers requires so-called super multiview displays with extremely high angular resolution [Takaki 2006; Takaki et al. 2011; Pamplona et al. 2012]; consequently, such displays remain limited in size and resolution. Recently, Maimone et al. [2013] proposed an architecture which uses a combination of a light-attenuating LCD stack and a high angular resolution backlight to steer light in the direction of the viewer, potentially supporting accommodation for human viewers. Head-mounted displays have a less severe angular resolution requirement; thus, the near-eye light field display of Lanman and Luebke [2013] is also capable of supporting correct accommodation.

The conceptual similarity between our work and the Layered 3D display of Wetzstein et al. [2011] bears some discussion. Although the physical realizations are quite different, both methods work with multiple display layers and use optimization to compute the displayed patterns so that the resulting imagery best matches specified views of the scene. However, the specifics of the problem differ significantly: their work deals with parallax, and consequently uses optimization techniques from tomographic reconstruction, while our goal is to support accommodation and defocus, for which we use an optimization technique designed for image processing applications. Nevertheless, it is interesting to observe that the optimal patterns produced by both methods appear qualitatively similar.

Volumetric displays place light sources in a three-dimensional volume, for example by using rotating display screens [Favalora et al. 2002] or stacks of switchable diffusers [Sullivan 2004]. These allow fully correct vergence, accommodation, and parallax, but the scene is restricted to the size of the display volume, and the large number of addressable voxels needed places practical limits on resolution. The biggest limitation of the above displays is that they present additive light, creating a scene composed of glowing, transparent voxels. This limitation makes it hard to reproduce occlusions and non-Lambertian effects. More recent techniques [Cossairt et al. 2007; Jones et al. 2007] have used anisotropic diffusers to overcome this limitation, but at a cost: accommodation cues become incorrect [Jones et al. 2007].

Multi-plane displays are a variation of volumetric displays in which the viewpoint is fixed. Such displays are very promising because they can in principle provide correct depth cues, including accommodation, with conventional display hardware. In multi-plane displays, images are drawn on presentation planes at several different optical distances for each eye, enabling both vergence and accommodation cues. These displays have been constructed using a system of beam splitters [Akeley et al. 2004; MacKenzie et al. 2010] and by time multiplexing with high-speed switchable lenses [Love et al. 2009; Liu et al. 2010] to superimpose multiple display planes additively on the viewer's field of vision (Figure 2). Many current implementations support high-resolution imagery using the full resolution of a conventional monitor, with accommodation cues that are correct for



Akeley et al. [2004] Love et al. [2009]

Figure 2: *Multi-plane display architectures. Akeley et al. [2004] use beam splitters to superimpose images of different parts of a monitor along the same viewing axis. Love et al. [2009] use high-speed switchable lenses to change the optical distance of the monitor; time multiplexing the image on the monitor in synchronization with the lenses creates the effect of several screen planes at different optical distances.*

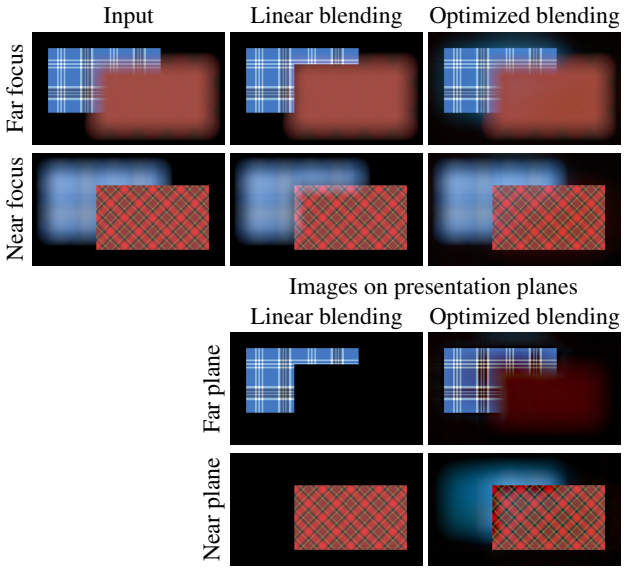


Figure 3: *Occlusion boundaries create noticeable artifacts in linear depth-weighted blending, whereas our optimization approach can render them accurately. Here, for illustration, we consider a two-plane display whose planes are aligned with the opaque rectangles in the input scene. Linear blending assigns intensities based on per-pixel distances in a pinhole image, creating spurious edges, darkening, and halos when defocus occurs. Our optimization algorithm, on the other hand, eliminates artifacts by automatically keeping the high-frequency content of the occlusion boundary on the near plane.*

diffuse objects lying on one of the presentation planes. Recent work by Hu and Hua [2013; 2014a; 2014b] has been building towards a high-resolution multi-plane head-mounted display that potentially supports a large number of focal planes.

MacKenzie et al. [2010] showed that viewers accommodate to the simulated distance of an object between planes when a per-pixel linear blending rule is used to simulate distances in between presentation planes. Unfortunately, per-pixel blending can only be used for simple diffuse scenes without occlusions. This class of three-dimensional displays has the potential to eliminate viewer discomfort, but it has not been clear so far how to reproduce general scenes with occlusions, reflections, and other non-Lambertian effects that do not exhibit a consistent accommodative distance.

In this work, we address this limitation of volumetric displays. Specifically, we present a general approach for reproducing arbitrary scenes on multi-plane displays. We note, however, that these displays do not reproduce parallax cues; furthermore, the viewer is constrained to a fixed location and cannot move freely relative to the display. For this reason, such displays are only practical for specialized applications at present, but could be incorporated into future head-mounted display designs such as those under development by Hu and Hua.

3 Optimal presentation of focus cues

Displaying a 3D scene on a nontraditional display architecture requires mapping its radiance distribution to the display elements of the device. Typically, this mapping is done with the goal of reproducing the scene appearance as closely as possible for an ideal viewer. For example, each ray of a traditional light field display typically displays the scene radiance sampled at the corresponding point in ray space after filtering [Zwicker et al. 2006]. Similarly, with a volumetric display, one may distribute the radiance of each scene point to its nearby display elements. In the case of a multi-plane display, an object at one of the presentation planes would be displayed on the pixels of that presentation plane, and objects between presentation planes would be displayed by distributing their intensities on the pixels of the adjacent presentation planes, whether by linear interpolation [Akeley et al. 2004; Hoffman et al. 2008; MacKenzie et al. 2010; Ryan et al. 2012] or more complex interpolation schemes [Liu and Hua 2010; Hu and Hua 2014a]. These per-pixel blending strategies work well for diffuse surfaces in scenes where depth varies slowly across the image, and give nearly correct accommodation cues for broadband stimuli [Ravikumar et al. 2011]. However, they also produces quite noticeable haloing artifacts around occlusions, as illustrated in Figure 3. Reflections, refractions, and other non-Lambertian phenomena also produce image features that cannot be assigned a consistent accommodative distance, and so cannot be handled with such methods. In general, it is not possible to reproduce accommodation cues by locally assigning depth to different components of the image.

We approach the problem from a different direction: our goal is to best reproduce the appearance of the scene as it would look to the viewer observing the scene directly, including changes in focus due to accommodation. Using a model of image formation in the eye, we can obtain for each desired viewpoint the focal stack of images that would be seen by the viewer when accommodating to different distances. For the purposes of display, we may treat this as a full description of the scene, because it encodes all the accommodation cues. Given this scene description, we optimize the assignment of light intensities to display elements so that the images seen in the display, again predicted with the image formation model, for all input accommodation distances are as close as possible to the images of the original scene.

In the following subsection, we apply this idea to the problem of presenting general scenes with accurate focus cues on multi-plane displays. Using this approach eliminates artifacts at occlusion boundaries, correctly handles specularity, and gives nearly correct focus cues in a variety of scenes including those acquired from the real world.

3.1 Focus cues on multi-plane displays

Our volumetric display system is based on that of Love et al. [2009], shown in Figure 2 (right). This is a multi-plane display that shows images at four additive presentation planes with the accommodative distances equally spaced in diopters (D), i.e. the reciprocal of the distance from the eye in meters. The separation between planes is $0.6\,\mathrm{D}$, yielding a full work space of $1.8\,\mathrm{D}$ that can be translated forwards or backwards in dioptric space by insertion of an additional lens before the eye. The setup is duplicated for each eye to provide stereoscopic imagery. Each eye can therefore be treated independently. Appopriate vergence and accommodation cues are provided through the volume.

For clarity, we will use the following notational conventions. Signals in the frequency domain are indicated with a hat; two signals denoted with the same letter, say u and $\hat{u}$, form a Fourier pair. Indexing into a signal will be denoted with square brackets, e.g. $u[x]$, to distinguish it from a family of signals parametrized by a variable, e.g. $u(t)$. Pointwise multiplication of signals is denoted by $u \cdot v$. The norm symbol denotes the L^2 norm, which for signals is $\|u\| = \left(\int u[x]^2 \,\mathrm{d}x\right)^{1/2}$. We assume that the Fourier transform is scaled such that $\|\hat{u}\| = \|u\|$.

Our approach relies on matching the defocus effects produced in the eye when accommodating at different distances. To do so, we require a model of image formation in the eye. While more sophisticated models of the human eye are available [Navarro 2009; Coelho et al. 2013], for our purposes the eye can be well approximated by a thin lens camera with a circular aperture. Human pupil diameter varies between 2 mm and 7 mm under ordinary circumstances [Spring and Stiles 1948]; we assume a diameter $a = 4\,\mathrm{mm}$ consistent with viewing images on a bright monitor.

Using the image formation model, we can predict the image $s(z)$ that would be seen by the viewer when accommodating to any specified distance z in the original scene. For a synthetic scene, we can easily compute $s(z)$ by rendering the scene with a virtual thin lens camera with aperture a focused to distance z. For a real-world scene, $s(z)$ corresponds to a photograph taken with a real camera with these parameters. (Note that it is in principle possible to adjust the optimization for a specific viewer using known optometric data, as long as appropriately defocused images of the desired scene can be obtained; however, doing so may only be practical for synthetic scenes.)

We can also predict the image seen when accommodating to the same distance when viewing the display itself. Suppose the n presentation planes of the display are located at distances of $z_1^{\mathrm{p}}, \ldots, z_n^{\mathrm{p}}$ and show images $p_1, \ldots, p_n$. Because the display planes are additive, the image $v(z)$ seen by the viewer when accommodating to distance z can be determined by adding up the contributions of each plane. The contribution of the jth plane is simply p_j convolved with the corresponding point spread function (PSF) of the eye, $h(z, z_j^{\mathrm{p}})$, which is given by the image formation model. The Fourier transform of the total image seen through the system is therefore given by

$$\hat{v}(z) = \sum_{j=1}^{n} \hat{h}(z, z_j^{\mathrm{p}}) \cdot \hat{p}_j. \tag{1}$$

Our goal then is to determine the intensities to present on the presentation planes, so that the image $v(z)$ seen through the display is close to the desired image of the scene $s(z)$ for all distances z in the range of interest. To do so, we require a metric for the error between the desired and the actual displayed images. The conventional least-squares approach, as used in much existing work on compressive displays [Wetzstein et al. 2011; Wetzstein et al. 2012b; Maimone et al. 2013], amounts to using the L^2 distance $\|s - v\|$ as the error metric. However, because the contrast sensitivity of the human visual system varies with frequency, it is more perceptually accurate to measure error taking this contrast sensitivity into account. Accordingly, we define the error as

$$E(s, v) = \|\hat{c} \cdot (\hat{s} - \hat{v})\|, \tag{2}$$

where $\hat{c}$ is the contrast sensitivity function, for which we use the model proposed by Mantiuk et al. [2011]. The traditional L^2 error corresponds to taking $\hat{c} = 1$ instead. As we shall see, an error metric defined in frequency space fits naturally into our optimization scheme. We currently ignore more complex perceptual effects such as signal-dependent masking and nonlinear contrast response.

We take several sample distances $z_1^{\mathrm{s}}, \ldots, z_m^{\mathrm{s}}$ uniformly spaced in diopters across the range of interest, and seek to minimize the total squared error

$$f(p_1, \ldots, p_n) = \sum_{i=1}^{m} E\big(s(z_i^{\mathrm{s}}), v(z_i^{\mathrm{s}})\big)^2 \tag{3}$$

$$= \sum_{i=1}^{m} \left\| \hat{c} \cdot \big(s(z_i^{\mathrm{s}}) - v(z_i^{\mathrm{s}})\big) \right\|^2. \tag{4}$$

Denoting $s(z_i^{\mathrm{s}})$ and $h(z_i^{\mathrm{s}}, z_j^{\mathrm{p}})$ by s_i and h_{ij} respectively, the objective can be written as

$$f(p_1, \ldots, p_n) = \sum_{i=1}^{m} \left\| \hat{c} \cdot \hat{s}_i - \sum_{j=1}^{n} \hat{c} \cdot \hat{h}_{ij} \cdot \hat{p}_j \right\|^2. \tag{5}$$

Because the entries of the p_j cannot exceed the device's displayable intensity range, say $[0, 1]$, our task is to solve the following minimization problem:

$$\begin{aligned} &\min f(p_1, \ldots, p_n) \\ &\text{s.t. } 0 \le p_j \le 1 \text{ for } j = 1, \ldots, n. \end{aligned} \tag{6}$$

Because the Fourier transform is linear, the objective f is a quadratic function of the display intensities, and can be written as

$$f(\mathsf{p}) = \|\mathsf{b} - \mathsf{A}\mathsf{p}\|^2, \tag{7}$$

where p is the concatenation of the display images p_j into a single vector. Thus the problem is an instance of bound-constrained quadratic programming. However, the system matrix $\mathsf{A}^T\mathsf{A}$ is extremely large and dense, and it is impractical to form the matrix explicitly. This fact remains true even for the simpler L^2 error ($\hat{c} = 1$), because A represents convolutions with the eye's PSFs at different focus distances, and the support of the PSFs increases quadratically with image resolution.

3.1.1 Related techniques

Problem (6) is closely related to certain problems studied in previous work. Because these previous techniques implicitly use an L^2 error metric due to their least-squares formulation, we will restrict our attention to the L^2 case in this subsection.

If we view the collection of unknown presentation images p_j as a volumetric signal to be recovered from its known linear projections s_i, this perspective leads us to tomographic reconstruction. Levoy et al. [2006] recovered a light field from a focus stack of microscope images using a variation of the SART tomographic reconstruction algorithm [Andersen and Kak 1984] that they adapted to the convolution case. This algorithm can be used directly for our problem, but

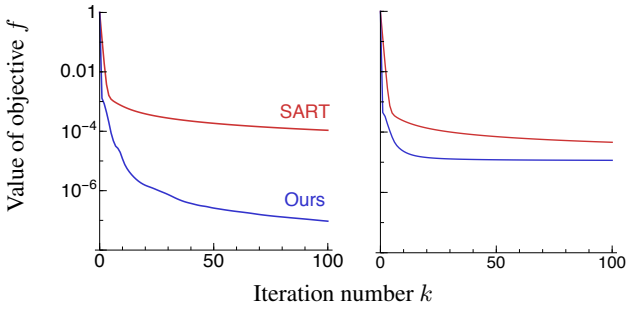


Figure 4: *A comparison of the convergence rates of our method and the SART algorithm (left) on an input that can be exactly represented on the display, and (right) on the input shown in Figure 10. We plot the value of f at the projection of p^k to the feasible set, relative to the initial value at $\mathsf{p}^0 = 0$. The objective f is defined using the L^2 error in all cases. Even when $f(\mathsf{p}) = 0$ cannot be attained, as on the right, our method converges much more quickly to the optimum.*

we found it to converge extremely slowly (Figure 4). The slow convergence may be because the method is a form of projected gradient descent, which tends to perform poorly when the system is poorly conditioned.

Another closely related technique is the Layered 3D architecture of Wetzstein et al. [2011], which uses a stack of multiplicative layers to reproduce a light field. In their system, each output ray is affected by only a few display elements; consequently, they obtain an extremely large but highly sparse problem. Our problem expressed in such a form would be smaller (though still large) but very dense, as each pixel is affected by a point-spread function which may span tens or hundreds of pixels. We use a frequency-domain approach to make this problem tractable, as described below.

3.2 Primal-dual optimization via the frequency domain

We have seen that the presence of convolutions renders the problem difficult to treat directly in terms of the unknown pixel values. In the frequency domain, however, all operations are pointwise, and so the objective is separable into a sum of squares for different spatial frequencies:

$$f = \sum_{i=1}^{m} \left\| \hat{c} \cdot \hat{s}_i - \sum_{j=1}^{n} \hat{c} \cdot \hat{h}_{ij} \cdot \hat{p}_j \right\|^2 \tag{8}$$

$$= \sum_{\xi} \sum_{i=1}^{m} \left| \hat{c}[\xi]\hat{s}_i[\xi] - \sum_{j=1}^{n} \hat{c}[\xi]\hat{h}_{ij}[\xi]\hat{p}_j[\xi] \right|^2 \tag{9}$$

$$= \sum_{\xi} \hat{c}[\xi]^2 \left\| \hat{\mathbf{s}}[\xi] - \hat{\mathbf{H}}[\xi]\hat{\mathbf{p}}[\xi] \right\|^2 \tag{10}$$

where ξ ranges over each of the frequency bases, and

$$\hat{\mathbf{s}}[\xi] = \begin{bmatrix} \hat{s}_1[\xi] & \dots & \hat{s}_m[\xi] \end{bmatrix}^T, \tag{11}$$

$$\hat{\mathbf{p}}[\xi] = \begin{bmatrix} \hat{p}_1[\xi] & \dots & \hat{p}_n[\xi] \end{bmatrix}^T, \tag{12}$$

$$\hat{\mathbf{H}}[\xi] = \begin{bmatrix} \hat{h}_{11}[\xi] & \dots & \hat{h}_{1n}[\xi] \\ \vdots & \ddots & \vdots \\ \hat{h}_{m1}[\xi] & \dots & \hat{h}_{mn}[\xi] \end{bmatrix}. \tag{13}$$

In the absence of constraints, the objective could easily be minimized by solving a small linear system for each spatial frequency. On the other hand, the constraints on pixel intensities live in the original image space and cannot be easily expressed in terms of frequency components. Therefore, it is natural to seek an optimization strategy that separates the objective and the constraints into primal and dual steps, allowing us to switch between image space and frequency space as appropriate. We use the primal-dual hybrid gradient (PDHG) algorithm [Zhu and Chan 2008; Esser et al. 2010], which can be seen as a preconditioned version of the alternating direction method of multipliers [Chambolle and Pock 2011]. The PDHG algorithm has been shown to perform remarkably well on large-scale optimization problems in image processing and computer vision, and we find that it converges efficiently on our problem as well. We detail its application to our problem below.

3.2.1 The PDHG algorithm

The PDHG algorithm applies to saddle point problems of the form

$$\min_{\mathsf{p}} \max_{\mathsf{q}} \; f(\mathsf{p}) + \langle \mathsf{K}\mathsf{p}, \mathsf{q} \rangle - g(\mathsf{q}), \tag{14}$$

where f and g are convex. A constrained convex problem can be put into this form by taking the indicator function of the feasible set, in our case

$$c(\mathsf{p}) = \begin{cases} 0 & \text{if } 0 \le \mathsf{p} \le 1, \\ \infty & \text{otherwise,} \end{cases} \tag{15}$$

and choosing g to be its convex conjugate,

$$g(\mathsf{q}) = \max_{\mathsf{p}} \langle \mathsf{p}, \mathsf{q} \rangle - c(\mathsf{p}) \tag{16}$$

$$= \sum_i \max(\mathsf{q}_i, 0) \tag{17}$$

defined over dual variables q of the same dimensionality as p. By duality, we also have $c(\mathsf{p}) = \max_{\mathsf{q}} \langle \mathsf{p}, \mathsf{q} \rangle - g(\mathsf{q})$. The original problem (6) is equivalent to minimizing $f(\mathsf{p}) + c(\mathsf{p})$, that is,

$$\min_{\mathsf{p}} \max_{\mathsf{q}} \; f(\mathsf{p}) + \langle \mathsf{p}, \mathsf{q} \rangle - g(\mathsf{q}), \tag{18}$$

which is the same as (14) with K being the identity.

The PDHG algorithm essentially consists of alternately iterating the following *primal* and *dual* steps:

$$\mathsf{p}^{k+1} = \arg\min_{\mathsf{p}} f(\mathsf{p}) + \langle \mathsf{q}^k, \mathsf{p} \rangle + \frac{1}{2\tau^k} \|\mathsf{p} - \mathsf{p}^k\|^2, \tag{19}$$

$$\mathsf{q}^{k+1} = \arg\min_{\mathsf{q}} g(\mathsf{q}) - \langle \mathsf{q}, \mathsf{p}^{k+1} \rangle + \frac{1}{2\sigma^k} \|\mathsf{q} - \mathsf{q}^k\|^2, \tag{20}$$

where $\sigma^k, \tau^k > 0$ are parameters that control the step size. For constant step sizes $\sigma^k = \sigma$, $\tau^k = \tau$, the algorithm converges as long as $\sigma\tau < 1$. Typically, the term $\langle \mathsf{q}, \mathsf{p}^{k+1} \rangle$ in the dual step is replaced with an overrelaxed term $\langle \mathsf{q}, 2\mathsf{p}^{k+1} - \mathsf{p}^k \rangle$, because this allows the largest step sizes with guaranteed convergence [He and Yuan 2012].

For the problem (6), the primal step can be expressed in frequency space as

$$\hat{\mathbf{p}}^{k+1} = \left(\hat{\mathbf{H}}^T \hat{\mathbf{H}} + \frac{1}{2\tau} \mathbf{I} \right)^{-1} \left(\hat{\mathbf{H}}^T \hat{\mathbf{s}} - \frac{1}{2} \hat{\mathbf{q}}^k + \frac{1}{2\tau} \hat{\mathbf{p}}^k \right) \tag{21}$$

for each frequency ξ (elided above for clarity). Here $\hat{\mathbf{p}}$, $\hat{\mathbf{s}}$, and $\hat{\mathbf{K}}$ are defined as in (11)-(13), and $\hat{\mathbf{q}}$ is defined analogously.

The minimization in the dual step is separable over the components of $\mathbf{q}$, each of which contributes a term

$$\max(q, 0) - q p^{k+1} + \frac{1}{2\sigma}(q - q^k)^2. \tag{22}$$

Algorithm 1 Computing the optimal presentation images using the PDHG algorithm.

function OPTIMIZE(scene images s, number of iterations N)
  Compute $\hat{\mathsf{s}}$ from s
  Initialize $\hat{\mathsf{p}}^0 = \mathsf{q}^0 = 0$
  for $k = 0, 1, \ldots, N-1$ **do**
    Compute $\hat{\mathsf{q}}^k$ from q^k
    Compute $\hat{\mathsf{p}}^{k+1}$ using (25)
    Compute p^{k+1} from $\hat{\mathsf{p}}^{k+1}$
    Compute q^{k+1} using (24)
  return $\mathrm{project}(\mathsf{p}^N, [0, 1])$

This is a piecewise quadratic which can be minimized analytically, yielding the update rule

$$\tilde{q} = q^k + \sigma p^{k+1}, \tag{23}$$

$$q^{k+1} = \begin{cases} \tilde{q} & \text{if } \tilde{q} < 0, \\ 0 & \text{if } 0 \le \tilde{q} \le \sigma, \\ \tilde{q} - \sigma & \text{if } \tilde{q} > \sigma. \end{cases} \tag{24}$$

For overrelaxed dual steps, define $\tilde{q} = q^k + \sigma(2p^{k+1} - p^k)$ instead.

We note two relevant implementation details. First, to avoid edge artifacts in the Fourier transform, we pad the input images s_i during initialization by replicating edge pixels and applying a smooth falloff to zero. Second, small amounts of noise in the input images cause spurious low-frequency oscillations in the result, because the problem is underdetermined for low frequencies. Ideally, this would have little effect on the final image because oscillations on different planes would cancel out, but in practice, the planes can be slightly misaligned, which renders these artifacts visible as "splotchiness" in the final image. We suppress these oscillations by adding a small amount of regularization to the primal step,

$$\hat{\mathbf{p}}^{k+1} = \left(\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \frac{1}{2\tau}\mathbf{I} + \frac{\epsilon}{\|\hat{\mathbf{s}}\|}\mathbf{I}\right)^{-1}\left(\hat{\mathbf{H}}^T\hat{\mathbf{s}} - \frac{1}{2}\hat{\mathbf{q}}^k + \frac{1}{2\tau}\hat{\mathbf{p}}^k\right) \tag{25}$$

with $\epsilon = 10^{-6}$. Making the regularization strength inversely proportional to signal amplitude causes only the frequencies containing small-amplitude noise to be damped out.

The full algorithm is given in pseudocode in Algorithm 1. We use $\tau = 10^2$ and $\sigma = 0.95/\tau$, which gave the fastest convergence in our experiments. Running the algorithm for a fixed number of iterations, $N = 100$, was sufficient for our results. We used a CPU-only implementation with parallelization through OpenMP. For 24 input images of size 600×600, the optimization takes about 3 minutes on a machine with a 4-core 3.5 GHz processor. The algorithm scales very well with resolution because all steps except the Fourier transform are linear in the number of pixels.

4 Analysis

Practical multi-plane displays have a relatively small number of presentation planes, so it is important to determine how many planes and what spacing are required to achieve specified goals. Fortunately, the ability of the human visual system to discriminate changes in focal distance is limited. Depth of focus—the change in focal distance that is just noticeable—is roughly 0.3 D under normal viewing conditions [Campbell 1957], so one could conceivably use that criterion as the basis for deciding how closely spaced presentation planes have to be.

Two lines of evidence suggest that the spacing can actually be quite a bit wider than that. First, accommodation can be driven essentially as well with a multi-plane display as with a real stimulus when the presentation planes are 0.9 D apart [MacKenzie et al. 2010; MacKenzie et al. 2012]. Second, accommodation and the perception of blur is determined primarily by spatial frequencies of 4 to 8 cpd [Owens 1980; Mathews and Kruger 1994; Granger and Cupery 1972], and at such frequencies plane separations of 0.6–0.7 D provide an excellent approximation to the real world (Figure 6; also Ravikumar et al. [2011]). Thus, with a 0.6 D spacing between planes it is possible to provide focus cues that are perceptually indistinguishable from the perfect case, for which we use the phrase "nearly correct focus cues".

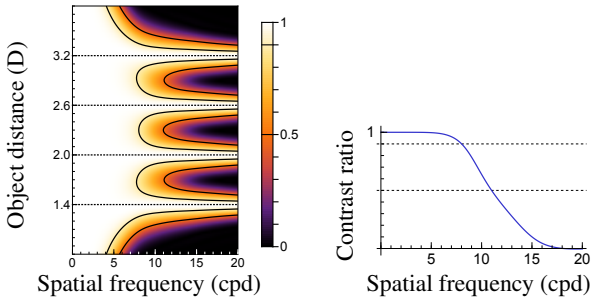


Figure 5: *Contrast ratio as a function of object distance and spatial frequency, when the viewer accommodates exactly to the simulated object. Contours are drawn at* 0.9 *and* 0.5*. The worst case occurs exactly between two presentation planes, for example at* $z^{-1} = 2.3\,\mathrm{D}$*; for this case, the variation with spatial frequency is plotted on the right.*

To better understand the properties and limitations of our approach, it is useful to look at the optimal solutions in certain simplified conditions. We also explore the theoretical limits of resolution and accommodative range of our display architecture. In this section, we assume that the pupil diameter is $a = 4\,\mathrm{mm}$, consistent with viewing images on a bright monitor, and presentation planes lie at dioptric distances $z^{\mathrm{p}}_{1,\ldots,4} = 1.4\,\mathrm{D}, 2.0\,\mathrm{D}, 2.6\,\mathrm{D}, 3.2\,\mathrm{D}$.

As a first approximation, we can neglect the constraints on displayable intensities if we assume the image content to consist of low amplitude variations on a constant background. Then the problem reduces to minimizing $\|\hat{\mathbf{s}} - \hat{\mathbf{H}}\hat{\mathbf{p}}\|^2$ for each frequency. Conceptually, we project $\hat{s}(z)$ to a linear combination of defocus responses $\hat{h}(z, z^{\mathrm{p}}_1), \ldots, \hat{h}(z, z^{\mathrm{p}}_n)$, and obtain $\hat{p}_1, \ldots, \hat{p}_n$ as the linear combination weights. Quantitatively, the matrix $\hat{\mathbf{H}}^T\hat{\mathbf{H}}$ determines the behavior of the solution. The matrix is close to diagonal for high spatial frequencies, but nearly rank 1 for low frequencies. In practical terms, this means that high-frequency image content must be assigned to its closest plane, while low frequencies are underconstrained and can be redistributed between planes. It is these underconstrained degrees of freedom that the algorithm exploits to satisfy the constraints on display intensities.

Due to the finite spacing of the presentation planes, there is a limit to the spatial resolution with which objects can be displayed at intermediate distances. Consider a scene consisting of a unit sinusoid of frequency ω at a distance $z \neq z^{\mathrm{s}}_1, \ldots, z^{\mathrm{s}}_n$; . When the viewer accommodates to z, the contrast seen in the displayed image will be different from the contrast in the original scene. In Figure 5, we plot the ratio between input contrast and retinal contrast for a range of distances between and beyond the presentation planes. The worst case occurs exactly halfway between two presentation planes; here, the contrast remains above 90% for up to 8 cpd, and reaches 50% at 11 cpd in the current setup. Better resolution, if needed, can be achieved by bringing the presentation planes closer together; the

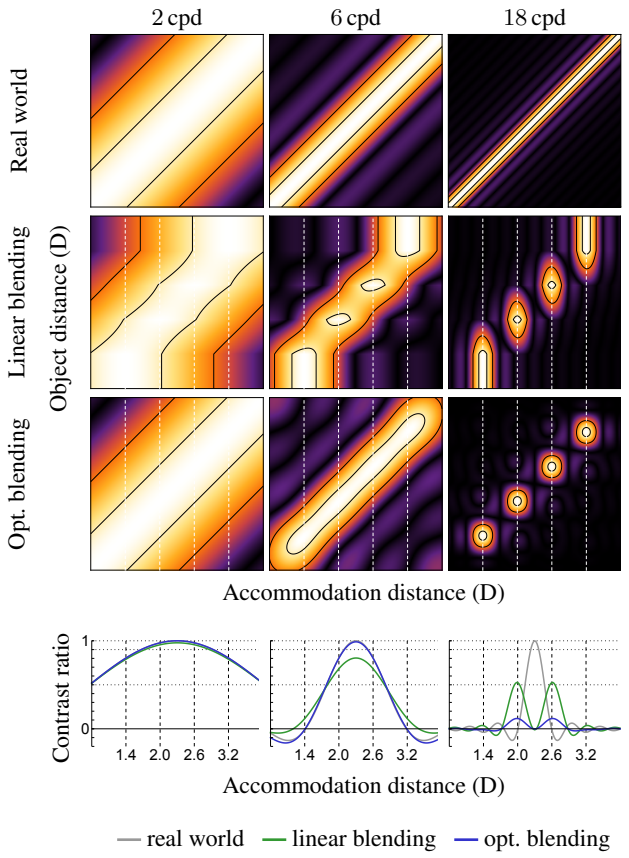


Figure 6: *Retinal contrast as a function of object distance and accommodation distance. The presentation planes are indicated with vertical dashed lines. Top: the behavior of real-world scenes, linear blending, and optimized blending is shown for spatial frequencies of* 2, 6, *and* 18 cpd. *Bottom: a horizontal slice through the middle of the plots, corresponding to an object at an intermediate distance of* 2.3 D. *Optimization retains higher contrast for moderate frequencies and projects undisplayable high frequencies to nearly zero. Linear blending creates spurious high-contrast peaks at the presentation planes, which may drive accommodation away from the simulated object's distance.*

threshold frequency is directly proportional to the reciprocal of the dioptric spacing between the planes.

Apart from the peak contrast of frequency components, which determines the maximum sharpness of the viewed image, we also consider how the contrast behaves when the viewer focuses at different distances, which determines the change in retinal-image contrast as accommodation varies. This is important because the gradient of image contrast drives the accommodative response of the eye, enabling the viewer to form a sharp image by accommodating in the direction of increasing contrast. In Figure 6, we show the retinal contrast of various spatial frequencies as a function of object distance and accommodation distance. The behavior of retinal contrast with our optimization method is closer to the ideal behavior than can be achieved with linear blending for all spatial frequencies.

For a qualitative visual evaluation, we set up a scene with a sinusoidal pattern with varying spatial frequency and contrast, placed at an intermediate distance of 2.3 D between two presentation planes (Figure 7). The variation in contrast also allows us to examine the effect of constraints on the optimal solution. The spatial frequency

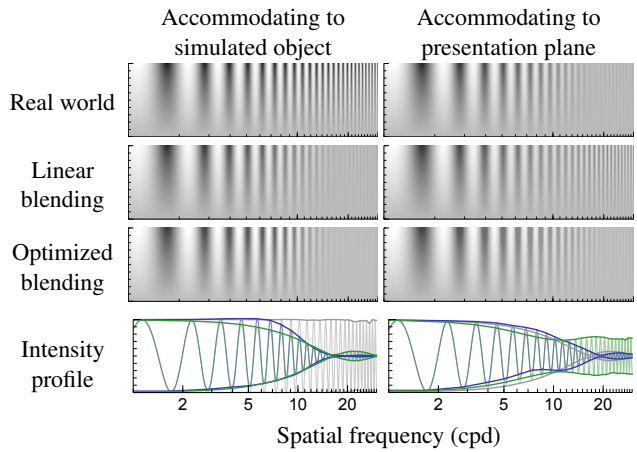


Figure 7: *Visual comparison of reproduction of spatial frequencies at an intermediate distance. The input is a sinusoidal pattern whose frequency increases horizontally from* 1 *to* 30 cpd *and whose contrast increases vertically. While neither method can reproduce high frequencies at this intermediate distance, linear blending incorrectly creates higher contrast at a different distance, as shown in the intensity profile taken at maximum contrast (cf. Figure 6).*

varies logarithmically from 1 to 30 cpd while its contrast varies independently from 0 to 1. As expected, higher spatial frequencies cannot be reproduced accurately at this intermediate distance. At higher contrast levels, we might also expect the reproduction to be worsened due to the constraints on the pixel intensities. A small amount of additional error can indeed be observed at medium frequencies in the error plot, but its magnitude remains below 0.08 up to 8 cpd, after which the unconstrained error begins to dominate. Linear blending is noticeably worse at high frequencies, creating a higher contrast when accommodating to the nearest presentation plane rather than to the desired distance of the pattern. This artificial increase in contrast may cause accommodation to be driven inappropriately to the presentation plane and not to the distance of the simulated object.

4.1 Light field analysis

To illuminate the differences between the additive multi-plane displays we consider and the multiplicative displays of recent work (e.g. Wetzstein et al. [2011], Maimone et al. [2013]), we briefly describe the characteristics and limitations of these displays in terms of light fields.

Because the planes in a multi-plane display span a very large range of optical distances, and because we want to analyze defocus effects occurring at the viewer's eye, we find it more convenient to adopt a viewer-centric light field parametrization based on view direction θ and translation across the pupil x, shown in Figure 8 (left). In this section, we will continue to use the term "spatial resolution" f_θ to refer to the directional resolution seen by the eye in cycles per degree, and to be consistent with previous papers we will refer to the translational resolution across the pupil as "angular resolution" f_x (although this leads to the slightly awkward situation that spatial and angular resolutions are expressed in units of cycles per degree and per mm respectively).

In Figure 8 (right) we plot the spatioangular frequency support of the multi-plane display used for our results along with that of the two displays compared in Maimone et al. [2013]. As each presentation plane of the multi-plane display is a diffuse emitter, its spectral

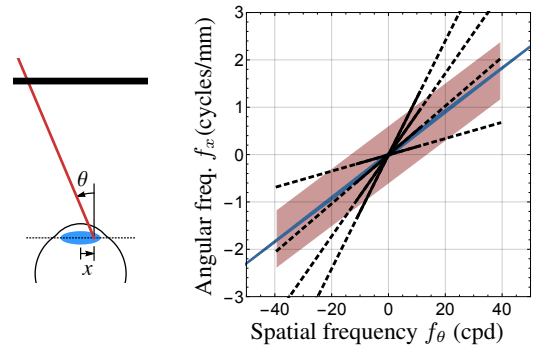


Figure 8: *Left: Our viewer-centric light field parametrization. Right: The spatioangular frequency support of various multilayer and multi-plane displays: a two-layer display with layer separation* 40 mm *(blue), the HAR backlight display of Maimone et al. [2013] (red), and our multi-plane display with presentation planes between* 0.3 D *and* 2.1 D *(solid black lines). The monitor we use to drive the display has a resolution of* 11 cpd*; changing to a higher-resolution monitor such as the one used by Maimone et al. [2013] would extend the support (dotted black lines).*

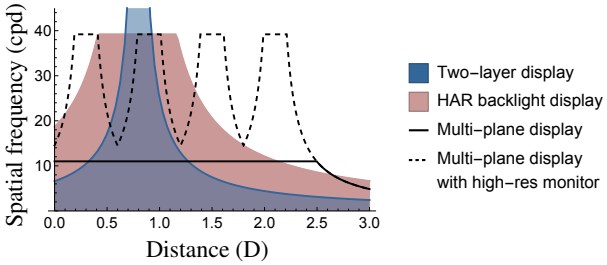


Figure 9: *Upper bounds on maximum displayable frequency when the viewer focuses at different distances. Aperture diameter is* $a = 4$ mm.

support is restricted to a line in spatioangular frequency space. In particular, a plane at distance d diopters corresponds to the line $f_x = \frac{180}{\pi} d f_\theta$ in our parametrization (due to the conversion between radians and degrees). Nevertheless, by using multiple planes at different distances, which show up as lines of different slopes, one can begin to sample a volume of the spatioangular frequency space.

To understand how the frequency support actually affects visual performance, we consider the image formed at the viewer when focusing to different distances. Focusing to a plane at dioptric distance d corresponds to convolving the light field spectrum along the f_x axis by the Fourier transform of the aperture, and taking the slice along the corresponding line $f_x = \frac{180}{\pi} d f_\theta$. Assuming the Fourier transform of an aperture of diameter a mm to be negligible beyond a frequency of $1/a$ cycles/mm, we can find an upper bound on the displayable resolution by finding the largest spatial frequency at which said line is within $1/a$ of the spatio-angular support of the display. This is a finite-aperture version of the "depth of field" as defined by Wetzstein et al. [2011]. We plot this maximum spatial frequency as a function of d in Figure 9. Note that the resolution of the monitor used in our implementation limits the spatial resolution at the eye to 11 cpd, so we also show the upper bound for a resolution equal to that of the configuration of Maimone et al. [2013], approximately 40 cpd. The multi-plane display is capable of a significantly wider depth of field than current multiplicative displays, although maximum resolution decreases at intermediate distances. Adding another switchable lens in the current display would yield eight independent presentation planes, thereby improving resolution without decreasing the depth of the workable volume.

5 Results

In this section, we show some more example scenes displayed using our algorithm. The imagery for these scenes is acquired from a variety of sources: renderings of a synthetic scene, a focus stack shot with a DSLR camera, and light fields from a Lytro camera. These results demonstrate the versatility of our approach and the ease of acquisition of the necessary source imagery. We point out, however, that the images shown here and in the supplemental video are all images with static focal distance, and do not capture the remarkable sense of realism that one experiences when one looks through the display and is able to focus at arbitrary distances. The perceptual studies described in 5.1 indicate that this subjective impression is real and accompanied by measurable effects.

The most obvious improvements with our algorithm compared to linear blending concern the defocus effects in the presence of occlusions, reflections, and refractions. While the differences are a bit subtle in printed images, they are very noticeable when viewed in the display. The artifacts produced by linear blending at occlusion boundaries are salient and distracting as the viewer accommodates to different distances. With our method occlusions yield much stronger impressions of depth. Furthermore, reflections and refractions can be displayed with appropriate defocus effects using our algorithm, but not with linear blending except in the special case of planar reflections. Curved specular surfaces give rise to reflections and refractions with astigmatic defocus, which do not correspond to an image at any single distance. Such non-Lambertian defocus effects do not pose a problem for our method.

In Figure 10, we show a synthetic scene containing diffuse, specular and refractive objects at various distances, rendered with the *Mitsuba* physically based renderer [Jakob 2010]. Using a thin lens camera model with a 4 mm diameter aperture, the view for each eye was rendered with the camera focused at distances of $0.0\,\text{D}, 0.1\,\text{D}, \ldots, 2.2\,\text{D}$. We computed optimal presentation images for four presentation planes with 0.6 D spacing between 0.2 D and 2.0 D. Using the image formation model, we can simulate the images that would be seen by the viewer when looking through the display system; these closely match the input images. Note that the algorithm automatically respects the optical distance of the images in the specular and refractive objects. For example, the image of the armadillo reflected in the specular mirror is sharp at far focus and blurred at near focus, unlike the mirror itself.

In the supplementary video, we show photographs of the system taken with a DSLR camera with a 4 mm aperture, validating the simulated images. We also show an animation of the scene with a moving camera, showing no temporal artifacts even though each frame is optimized independently. Such animations can be readily displayed on our physical system at up to 45 fps.

We used the visual metric HDR-VDP-2 [Mantiuk et al. 2011] to quantify the perceptual impact of our technique. We compared the results of optimization and of linear blending to the ideal images for the scene in Figure 10. The per-pixel detection probability is visualized in Figure 11. With optimization, the maximum detection probability varies between 12.4% and 22.8% for different accommodation distances; with linear blending, it is 100% for all distances.

It is straightforward to use our approach to reproduce the defocus and accommodation cues associated with a real-world scene, because the algorithm does not require any knowledge of scene geometry, nor a full 4D light field. It is only necessary to acquire a series of images taken with the lens focused at known distances in the scene.

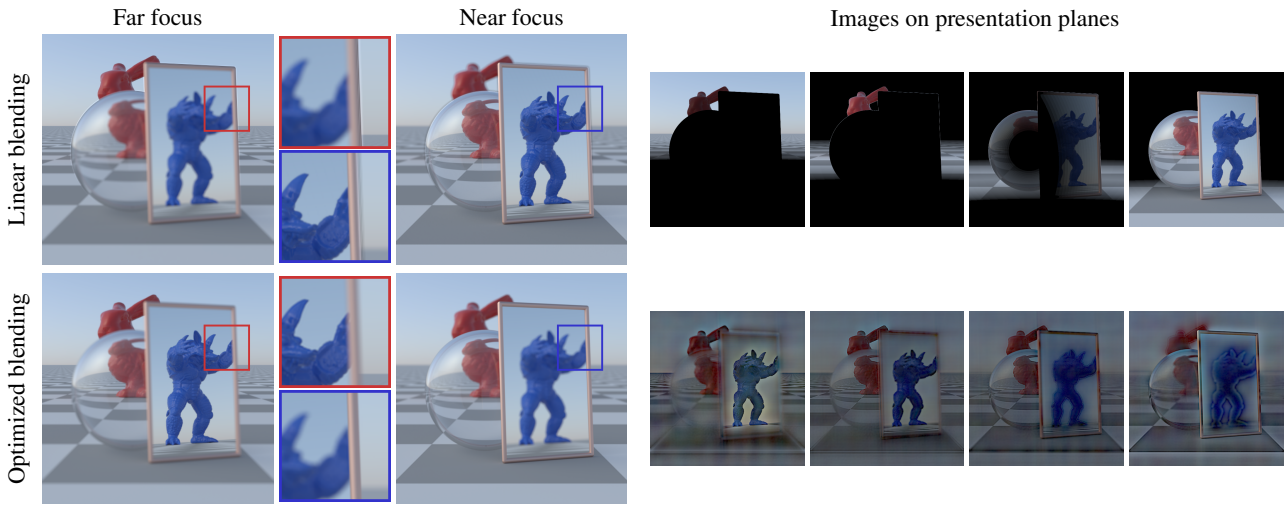


Figure 10: *Simulated images of a synthetic scene mapped to a multi-plane display. Reflections and refractions, such as the armadillo seen in the specular rectangle, produce stimuli which have an optical distance different from the distance of the surface. The optimization algorithm handles all these cases gracefully. Linear blending, on the other hand, produces severe artifacts such as sharp silhouettes of defocused objects, halos at depth discontinuities, and incorrect accommodation cues for reflections and refractions. For the plane mirror, the image could be assigned to the reflected depth in linear blending, but it is not possible to do so in general for curved objects such as the refractive sphere.*

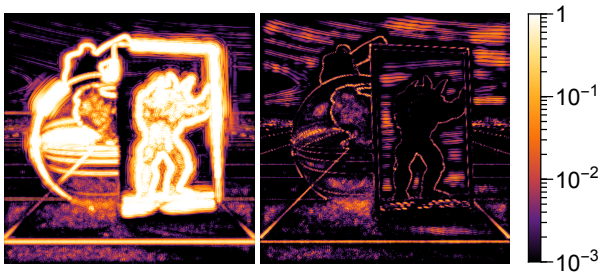


Figure 11: *Reproduction quality as measured by HDR-VDP-2. Images show the per-pixel probability of detecting a difference between the input scene and the scene viewed through the display, taking the maximum over all accommodation distances.*

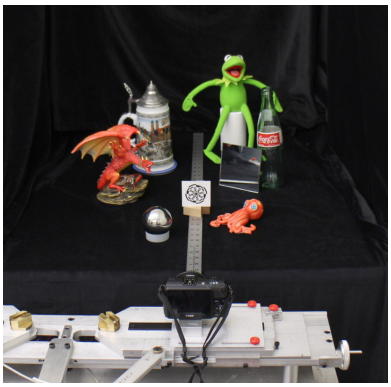


Figure 12: *Acquisition setup for the scene in Figure 13.*

We did so by placing a steel ruler in the scene parallel to the optical axis of the camera, with the zero marker aligned with the image plane, and moving a focus target to chosen distances along the ruler (Figure 12). Thus, with the aperture of the camera set to 4 mm, we acquired a series of images with known focal distances. Stereo imagery was obtained by translating the camera horizontally and repeating the process. Camera lenses typically exhibit "breathing", i.e. a change in magnification upon refocusing, which we removed in a postprocess by manual registration. After this procedure, we have a focal stack of images s_i focused at known distances z_i^{s}, on which we run our algorithm to obtain the result shown in Figure 13.

Light fields, being a source of refocusable imagery, can also be used as input to our algorithm. However, to do so requires a light field with sufficiently high angular resolution that it can resolve defocus effects with small changes in focus distance. We have used the first-generation Lytro light field camera to acquire such imagery. Unfortunately, the blur quality recorded by the camera for small amounts of defocus turns out to be insufficient to reproduce accommodation cues with our method. We obtained better results by capturing scenes with much larger dioptric ranges and compressing their nominal depth down to the 1.8 D range of the display. Photographed images of these scenes presented on the display are shown in Figures 1 and 14. While this procedure does not result in a physically accurate reproduction of the scene, we present it as a proof of concept. As commercially available light field cameras such as the Lytro ILLUM increase in resolution, we anticipate that in the future it will be possible to use their images without modification, and thus easily acquire and display high-quality imagery of real-world scenes with accurate accommodation cues.

5.1 Experiments

In vision science, the study of depth perception relies on experiments that display and manipulate numerous depth cues including disparity, perspective, shading, and more. Unfortunately, this effort has been hampered by an inability to present appropriate focus cues. In fact, some apparent misperceptions of depth have been shown to be the result of presenting inappropriate focus cues (and perhaps other naturally occurring cues) [Buckley and Frisby 1993; van Ee et al. 1999; Watt et al. 2005]. Thus, the ability to present appropriate cues with precise experimental control will be very useful to this area of

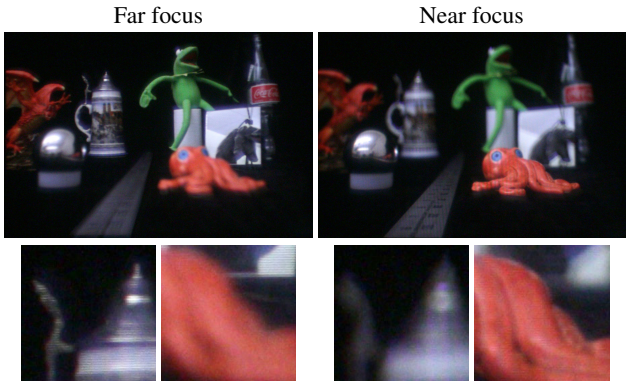


Figure 13: *Stereo imagery acquired from a real-world scene using a conventional DSLR camera (Figure 12) and presented on the multi-plane display.*

scientific research.

Our optimization technique has been employed in multiple psychophysical experiments testing the human visual system's response to accommodation cues from occlusions and reflections. Such experiments had not been possible before because occlusions and reflections could not be accurately reproduced by linear blending, the existing standard approach in this area. We briefly describe the experiments here to further validate the efficacy of the optimization technique, and to illustrate the importance of focus cues in visual perception.

One experiment, reported by Zannoli et al. [2014], examined peoples' ability to determine which of two surfaces is nearer when one occludes the other. Marshall et al. [1996] had pointed out that the blur of the boundary indicates which of the two surfaces is nearer: when the boundary is blurred, the blurrier surface is the occluder; when the boundary is sharp, the sharper surface is the occluder. Despite this completely informative cue, Marshall et al. found that people frequently perceive the blurred surface as near whether the boundary is blurred or sharp. The same result was obtained in two other psychophysical experiments [Mather and Smith 2002; Palmer and Brooks 2008]. The stimuli in all three experiments were rendered for and displayed on a single plane. Zannoli et al. replicated their results for single-plane stimuli: most people said the blurrier surface was near regardless of the blur of the boundary. The same occlusion relationship was then displayed in a multi-plane display with stimuli rendered with our optimization algorithm. Performance improved significantly in all subjects. That is, they generally perceived the physically nearer surface as nearer, even when viewing the stimuli monocularly, with no motion parallax, at stimulus durations that were too brief for an accommodative change to have occurred. This result shows that the multi-plane display with optimized rendering significantly improves viewers' ability to perceive the correct depth order at an occlusion boundary.

Another experiment, reported by Banks et al. [2014], tested the usefulness of the optimization technique and multi-plane display in presenting specular materials. Specular materials were presented volumetrically with optimized blending or non-volumetrically with a simulated aperture that matched the diameter of the viewer's pupil. Subjects perceived the volumetric stimuli as much glossier than the non-volumetric stimuli. This result shows that the multi-plane display with optimized blending yields more convincing impression of glossy material.

Figure 14: *Real-world scenes acquired with a Lytro camera and displayed on a multi-plane display.*

6 Conclusion

Focus and accommodation are subtle but important depth cues that have so far been neglected in most 3D displays. Multi-plane displays have been shown to present high-resolution imagery with nearly correct focus cues in simple cases; our technique greatly extends their capability by allowing arbitrary scenes with occlusions, reflections, and other non-Lambertian phenomena to be displayed with realistic focus cues.

In this paper, our goal has been to define the optimal solution to the problem of displaying arbitrary scenes on multi-plane displays, and compute it in a tractable way. As such, the algorithm we have presented converges much more quickly than existing tomographic techniques, but is still far from being fast enough for real-time applications. We hope that our characterization of the solution will inspire future work that addresses real-time performance.

Our algorithm may also be seen as a specific instantiation of a general approach for improving the presentation of accommodation cues in nontraditional display architectures through optimization. In particular, rather than driving display outputs so that they correspond directly to sampled values from the desired scene, the outputs can instead be set to the optimized values that would best create the desired perception. For example, we believe that our approach could be applied to other near-eye display architectures. The work of Lanman and Luebke [2013] resolves several views over the extent of the pupil, but each view is sampled as a pinhole camera. As a result, for any given accommodation distance the display will present sharper detail than can seen with a human eye. By taking into account the defocus in the eye one could potentially use that excess capacity to suppress artifacts that currently occur with direct sampling.

This work is part of a set of recent papers that directly address the vergence-accommodation conflict. These form one of the first steps in computer graphics towards practical 3D displays that support correct accommodation. Ultimately, these techniques will enable displays of the future to support natural viewing of 3D scenes with a degree of realism approaching that of the real world.

Acknowledgements

We thank Doug Lanman, Kurt Akeley, and the members of the Berkeley Visual Computing Group and the Banks Lab for their help, and the anonymous reviewers for their insightful comments. This work was supported by NSF awards BCS-1354029, CNS-1444840, and IIS-1353155, NEI training grant T32EY007043, and gifts from Intel, Pixar, and Qualcomm.

References

AKELEY, K., WATT, S. J., GIRSHICK, A. R., AND BANKS, M. S. 2004. A stereo display prototype with multiple focal distances. *ACM Trans. Graph. 23*, 3 (Aug.), 804–813.

ANDERSEN, A., AND KAK, A. 1984. Simultaneous algebraic reconstruction technique (SART): A superior implementation of the ART algorithm. *Ultrasonic Imaging 6*, 1, 81 – 94.

BANKS, M. S., BULBUL, A., ALBERT, R. A., NARAIN, R., O'BRIEN, J. F., AND WARD, G. 2014. The perception of surface material from disparity and focus cues. In *Proc. Vision Sciences Society 14th Annual Meeting*.

BUCKLEY, D., AND FRISBY, J. P. 1993. Interaction of stereo, texture and outline cues in the shape perception of three-dimensional ridges. *Vision Research 33*, 7, 919 – 933.

CAMPBELL, F. 1957. The depth of field of the human eye. *Optica Acta: International Journal of Optics 4*, 4, 157–164.

CHAMBOLLE, A., AND POCK, T. 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision 40*, 1, 120–145.

COELHO, J. M. P., BAIÃO, A., AND VIEIRA, P. 2013. Development of an optical simulator of the human eye. *Proc. SPIE 8785*, 8785CS–8785CS–8.

COSSAIRT, O. S., NAPOLI, J., HILL, S. L., DORVAL, R. K., AND FAVALORA, G. E. 2007. Occlusion-capable multiview volumetric three-dimensional display. *Appl. Opt. 46*, 8 (Mar), 1244–1250.

DU, S.-P., MASIA, B., HU, S.-M., AND GUTIERREZ, D. 2013. A metric of visual comfort for stereoscopic motion. *ACM Trans. Graph. 32*, 6 (Nov.), 222:1–222:9.

EMOTO, M., NIIDA, T., AND OKANO, F. 2005. Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television. *Display Technology, Journal of 1*, 2, 328–340.

ESSER, E., ZHANG, X., AND CHAN, T. 2010. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences 3*, 4, 1015–1046.

FAVALORA, G. E., NAPOLI, J., HALL, D. M., DORVAL, R. K., GIOVINCO, M., RICHMOND, M. J., AND CHUN, W. S. 2002. 100-million-voxel volumetric display. In *Proc. SPIE*, vol. 4712, 300–312.

GRANGER, E., AND CUPERY, K. 1972. Optical merit function (SQF), which correlates with subjective image judgments. *Photographic Science and Engineering 16*, 3.

HE, B., AND YUAN, X. 2012. Convergence analysis of primal-dual algorithms for a saddle-point problem: From contraction perspective. *SIAM Journal on Imaging Sciences 5*, 1, 119–149.

HELD, R. T., COOPER, E. A., O'BRIEN, J. F., AND BANKS, M. S. 2010. Using blur to affect perceived distance and size. *ACM Transactions on Graphics 29*, 2 (Mar.), 19:1–16.

HIRSCH, M., AND LANMAN, D., 2010. Build your own 3D display. ACM SIGGRAPH ASIA Course Notes.

HOFFMAN, D. M., GIRSHICK, A. R., AKELEY, K., AND BANKS, M. S. 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision 8*, 3.

HU, X., AND HUA, H. 2013. An optical see-through multi-focal-plane stereoscopic display prototype enabling nearly correct focus cues. vol. 8648, 86481A.

HU, X., AND HUA, H. 2014. Design and assessment of a depth-fused multi-focal-plane display prototype. *J. Display Technol. 10*, 4 (Apr), 308–316.

HU, X., AND HUA, H. 2014. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. *Opt. Express 22*, 11 (Jun), 13896–13903.

HUANG, F.-C., LANMAN, D., BARSKY, B. A., AND RASKAR, R. 2012. Correcting for optical aberrations using multilayer displays. *ACM Transaction on Graphics 31* (Nov.).

HUANG, F.-C., WETZSTEIN, G., BARSKY, B. A., AND RASKAR, R. 2014. Eyeglasses-free display: Towards correcting visual aberrations with computational light field displays. *ACM Trans. Graph. 33*, 4 (July), 59:1–59:12.

IVES, F., 1903. Parallax stereogram and process of making same., Apr. 14. US Patent 725,567.

JAKOB, W., 2010. Mitsuba renderer. http://www.mitsuba-renderer.org.

JONES, A., MCDOWALL, I., YAMADA, H., BOLAS, M., AND DEBEVEC, P. 2007. Rendering for an interactive 360° light field display. *ACM Trans. Graph. 26*, 3 (July).

LAMBOOIJ, M., FORTUIN, M., HEYNDERICKX, I., AND IJSSELSTEIJN, W. 2009. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of Imaging Science and Technology 53*, 3, 30201–1–30201–14.

LANG, M., HORNUNG, A., WANG, O., POULAKOS, S., SMOLIC, A., AND GROSS, M. 2010. Nonlinear disparity mapping for stereoscopic 3D. *ACM Trans. Graph. 29*, 4 (July), 75:1–75:10.

LANMAN, D., AND LUEBKE, D. 2013. Near-eye light field displays. *ACM Trans. Graph. 32*, 6 (Nov.), 220:1–220:10.

LANMAN, D., HIRSCH, M., KIM, Y., AND RASKAR, R. 2010. Content-adaptive parallax barriers: Optimizing dual-layer 3D displays using low-rank light field factorization. *ACM Trans. Graph. 29*, 6 (Dec.), 163:1–163:10.

LANMAN, D., WETZSTEIN, G., HIRSCH, M., HEIDRICH, W., AND RASKAR, R. 2011. Polarization fields: Dynamic light field display using multi-layer LCDs. *ACM Trans. Graph. 30*, 6.

LEVOY, M., NG, R., ADAMS, A., FOOTER, M., AND HOROWITZ, M. 2006. Light field microscopy. *ACM Trans. Graph. 25*, 3 (July), 924–934.

LIPPMANN, G. 1908. Épreuves réversibles donnant la sensation du relief. *J. Phys. Theor. Appl. 7*, 1, 821–825.

LIU, S., AND HUA, H. 2010. A systematic method for designing depth-fused multi-focal plane three-dimensional displays. *Opt. Express 18*, 11 (May), 11562–11573.

LIU, S., HUA, H., AND CHENG, D. 2010. A novel prototype for an optical see-through head-mounted display with addressable focus cues. *IEEE Transactions on Visualization and Computer Graphics 16*, 3, 381–393.

LOVE, G. D., HOFFMAN, D. M., HANDS, P. J., GAO, J., KIRBY, A. K., AND BANKS, M. S. 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. *Opt. Express 17*, 18 (Aug), 15716–15725.

MACKENZIE, K. J., HOFFMAN, D. M., AND WATT, S. J. 2010. Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control. *Journal of Vision 10*, 8.

MACKENZIE, K., DICKSON, R., AND WATT, S. 2012. Vergence and accommodation to multiple-image-plane stereoscopic displays: "real world" responses with practical image-plane separations? *Journal of Electronic Imaging 21*, 1.

MAIMONE, A., WETZSTEIN, G., LANMAN, D., HIRSCH, M., RASKAR, R., AND FUCHS, H. 2013. Focus 3D: Compressive accommodation display. *ACM Trans. Graph. 32*, 5, 1–13.

MANTIUK, R., KIM, K. J., REMPEL, A. G., AND HEIDRICH, W. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph. 30*, 4 (July), 40:1–40:14.

MARSHALL, J. A., BURBECK, C. A., ARIELY, D., ROLLAND, J. P., AND MARTIN, K. E. 1996. Occlusion edge blur: a cue to relative visual depth. *Journal of the Optical Society of America. A, Optics, image science, and vision 13*, 4 (Apr.), 681–8.

MASIA, B., WETZSTEIN, G., DIDYK, P., AND GUTIERREZ, D. 2013. A survey on computational displays: Pushing the boundaries of optics, computation, and perception. *Computers and Graphics 37*, 8, 1012 – 1038.

MATHER, G., AND SMITH, D. R. R. 2002. Blur discrimination and its relation to blur-mediated depth perception. *Perception 31*, 10, 1211–1219.

MATHEWS, S., AND KRUGER, P. 1994. Spatiotemporal transfer function of human accommodation. *Vision Research 34*, 15.

MATUSIK, W., AND PFISTER, H. 2004. 3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Trans. Graph. 23*, 3 (Aug.), 814–824.

MENDIBURU, B. 2009. *3D movie making: Stereoscopic digital cinema from script to screen*. Focal Press, Elsevier.

NAVARRO, R. 2009. The optical design of the human eye: a critical review. *Journal of Optometry 2*, 1, 3 – 18.

OWENS, D. 1980. A comparison of accommodative responsiveness and contrast sensitivity for sinusoidal gratings. *Vision Research 20*, 2, 159–167.

PALMER, S. E., AND BROOKS, J. L. 2008. Edge-region grouping in figure-ground organization and depth perception. *Journal of Experimental Psychology: Human Perception and Performance 34*, 6 (Dec), 1353–1371.

PAMPLONA, V. F., OLIVEIRA, M. M., ALIAGA, D. G., AND RASKAR, R. 2012. Tailored displays to compensate for visual aberrations. *ACM Trans. Graph. 31*, 4 (July), 81:1–81:12.

PERLIN, K., PAXIA, S., AND KOLLIN, J. S. 2000. An autostereoscopic display. In *Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, 319–326.

RAVIKUMAR, S., AKELEY, K., AND BANKS, M. S. 2011. Creating effective focus cues in multi-plane 3D displays. *Opt. Express 19*, 21 (Oct), 20940–20952.

RYAN, L., MACKENZIE, K., AND WATT, S. 2012. Multiple-focal-planes 3D displays: A practical solution to the vergence-accommodation conflict? In *3D Imaging (IC3D), 2012 International Conference on*, 1–6.

SHIBATA, T., KIM, J., HOFFMAN, D. M., AND BANKS, M. S. 2011. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision 11*, 8.

SPRING, K., AND STILES, W. S. 1948. Variation of pupil size with change in the angle at which the light stimulus strikes the retina. *British J. Ophthalmol. 32*, 6, 340–346.

SULLIVAN, A. 2004. DepthCube solid-state 3D volumetric display. In *Proc. SPIE*, vol. 5291, 279–284.

TAKAKI, Y., TANAKA, K., AND NAKAMURA, J. 2011. Super multi-view display with a lower resolution flat-panel display. *Opt. Express 19*, 5 (Feb), 4129–4139.

TAKAKI, Y. 2006. High-density directional display for generating natural three-dimensional images. *Proc. IEEE 94*, 3, 654–663.

VAN EE, R., BANKS, M. S., AND T., B. B. 1999. An analysis of binocular slant contrast. *Perception 28*, 9, 1121–1145.

WATT, S. J., AKELEY, K., ERNST, M. O., AND BANKS, M. S. 2005. Focus cues affect perceived depth. *Journal of Vision 5*, 10.

WETZSTEIN, G., LANMAN, D., HEIDRICH, W., AND RASKAR, R. 2011. Layered 3D: Tomographic image synthesis for attenuation-based light field and high dynamic range displays. *ACM Trans. Graph. 30*, 4.

WETZSTEIN, G., LANMAN, D., GUTIERREZ, D., AND HIRSCH, M., 2012. Computational displays: Combining optical fabrication, computational processing, and perceptual tricks to build the displays of the future. ACM SIGGRAPH Course Notes.

WETZSTEIN, G., LANMAN, D., HIRSCH, M., AND RASKAR, R. 2012. Tensor displays: Compressive light field synthesis using multilayer displays with directional backlighting. *ACM Trans. Graph. (Proc. SIGGRAPH) 31*, 4, 1–11.

ZANNOLI, M., ALBERT, R. A., BULBUL, A., NARAIN, R., O'BRIEN, J. F., AND BANKS, M. S. 2014. Correct blur and accommodation information is a reliable cue to depth ordering. In *Proc. Vision Sciences Society 14th Annual Meeting*.

ZHU, M., AND CHAN, T. 2008. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. Tech. rep., University of California, Los Angeles.

ZWICKER, M., MATUSIK, W., DURAND, F., PFISTER, H., AND FORLINES, C. 2006. Antialiasing for automultiscopic 3D displays. In *ACM SIGGRAPH 2006 Sketches*, SIGGRAPH '06.