Towards Multifocal Displays with Dense Focal Stacks

Supplemental Material

JEN-HAO RICK CHANG, Carnegie Mellon University, USA
B. V. K. VIJAYA KUMAR, Carnegie Mellon University, USA
ASWIN C. SANKARANARAYANAN, Carnegie Mellon University, USA

A LIGHT FIELD ANALYSIS

This section provides a detailed derviation of the analysis discussed in Section 3 of the main paper in detail. This analysis follows closely to the one in [Narain et al. 2015]. A notable difference however is that we provide analytical expressions for the perceived spatial resolution (Equation (3) in the main paper) and the minimum number of focal planes required (Equation (5)), whereas they only provide numerical results. For simplicity, we consider a flatland where a light field is two-dimensional and is parameterized by intercepts with two parallel axes, x and u. The two axes are separated by 1 unit, and for each x, we align the origin of u-axis to x. We model the human eye with a camera model that is composed of a finite-aperture lens and a sensor plane d_e away from the lens, as that used by Mercier et al. [2017] and Sun et al. [2017] We assume that the display and the sensor emits and receives light isotropically so that each pixel on the display uniformly emits light rays toward every direction, and vice versa for the sensor.

Light Field Generated by a Display. Let us decompose the optical path from the display to the retina (sensor) and examine the effect in frequency domain due to each component. Due to the finite pixel pitch, the light field creates by the display can be model as

$$\ell_d(x,u) = \left(\text{rect}\left(\frac{x}{\Delta x}\right) * \ell_t(x, u=0)\right) \times \sum_{m=-\infty}^{\infty} \delta(x - m\Delta x),$$

where $*$ represents two-dimensional convolution, Δx is the pitch of the display pixel, and ℓ_t is the target light field. The Fourier transform of $\ell_d(x,u)$ is

$$L_d(f_x,f_u) = (\text{sinc}(\Delta x f_x)\ \delta(f_u)\ L_t(f_x,f_u)) * \sum_{m=-\infty}^{\infty} \delta(f_x - \frac{m}{\Delta x}).$$

The finite pixel pitch acts as an anti-aliasing filter and thus we consider only the central spectrum replica ($m = 0$). Also, we assume $|L_t(f_x,f_u)| = 0$ for all $|f_x| \geq \frac{1}{2\Delta x}$ to avoid aliasing. Since the light field is nonnegative, or $\ell_d \geq 0$, we have $|L_t(f_x,f_u)| \leq L_t(0,0)$. Therefore, we have

$$|L_d(f_x,f_u)| \leq L_t(0,0)\ |\text{sinc}(\Delta x f_x)|\ \delta(f_u), \quad |f_x| \leq \frac{1}{2\Delta x} \tag{1}$$

$$|L_d(f_x,f_u)| = 0, \quad \text{otherwise.} \tag{2}$$

Authors' addresses: Jen-Hao Rick Chang, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213, USA, rickchang@cmu.edu; B. V. K. Vijaya Kumar, Carnegie Mellon University, Pittsburgh, USA, kumar@ece.cmu.edu; Aswin C. Sankaranarayanan, Carnegie Mellon University, Pittsburgh, USA, saswin@andrew.cmu.edu.

Therefore, in the ensuing derivation, we will focus on the upper-bound

$$\widehat{L}_d = \text{sinc}(\Delta x f_x)\ \delta(f_u)\ \text{rect}\left(\frac{f_x}{\Delta x}\right).$$

The light field spectrum $\widehat{L}_d$ forms a line segment parallel to f_x, as plotted in Figure 1a.

Propagation to the eye. After leaving the display, the light field propagates d_o and get refracted by the focus-tunable lens before reaching the eye. Under first-order optics, there operations can be modeled by coordinate transformation of the light fields [Hecht 2002]. Let $\mathbf{x} = [x\ u]^\top$. After propagating a distance d_o, the output light field is a reparameterization of the input light field and can be represented as

$$\ell_o(\mathbf{x}) = \ell_i\left(P_{d_o}^{-1}\mathbf{x}\right), \text{where } P_{d_o} = \begin{bmatrix} 1 & d_o \\ 0 & 1 \end{bmatrix}.$$

After refracted by a thin lens with focal length f, the output light field right after the lens is

$$\ell_o(\mathbf{x}) = \ell_i\left(R_f^{-1}\mathbf{x}\right), \text{where } R_f = \begin{bmatrix} 1 & 0 \\ \frac{-1}{f} & 1 \end{bmatrix}.$$

Since P_{d_o} and R_f are invertible, we can use the stretch theorem of d-dimensional Fourier transform to analyze their effect in the frequency domain. The general stretch theorem states that: Let $\mathbf{x} \in \mathbb{R}^d$, $\mathcal{F}(\cdot)$ be the Fourier transform operator, and $A \in \mathbb{R}^{d\times d}$ be any invertible matrix. We have

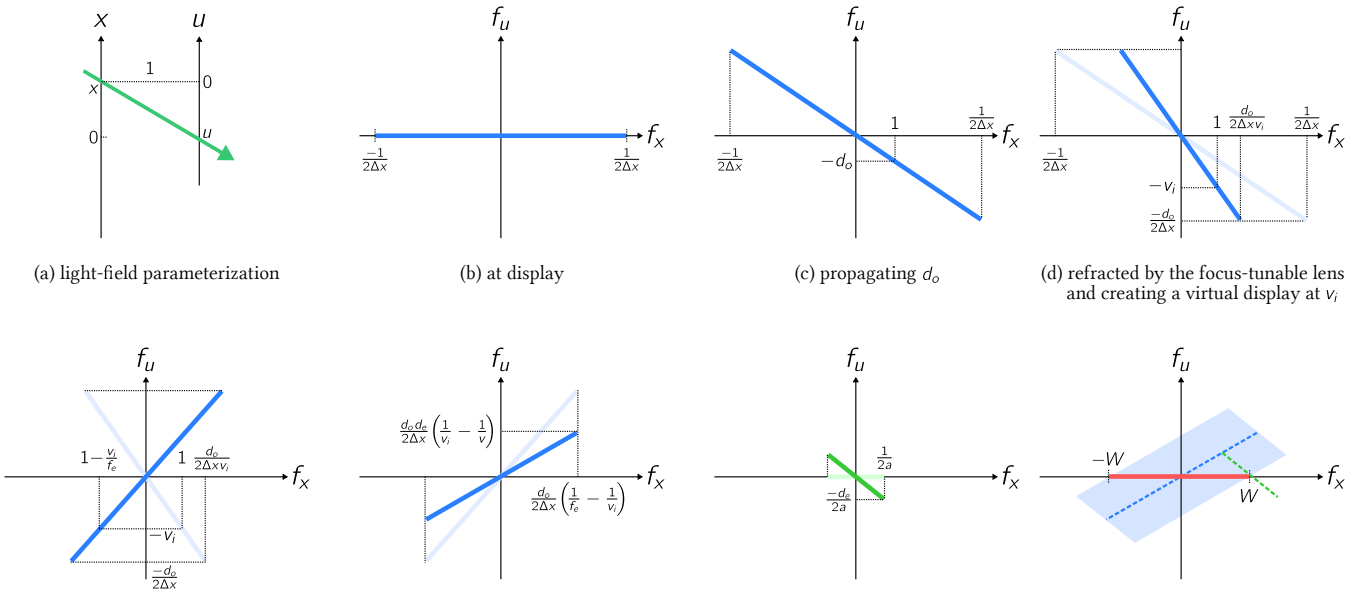
$$\mathcal{F}\left(\ell(A\mathbf{x})\right) = \frac{1}{|\det A|} L(A^{-\top}\mathbf{f}),$$

where L is the Fourier transform of ℓ, $\mathbf{f} \in \mathbb{R}^d$ is the variable in frequency domain, det A represents determinant of A, and $A^{-\top} = \left(A^\top\right)^{-1} = \left(A^{-1}\right)^\top$. By applying the stretch theorem to P_{d_o} and R_f, we can see that propagation and refraction shears the Fourier transform of the light field along f_u and f_x, respectively, as shown in Figure 1c-d.

Light Field Incident on the Retina. After reaching the eye, the light field ℓ_o is partially blocked by the pupil, refracted by the lens of the eye, propagates d_e to the retina, and finally integrated through all directions to form an image. The light field reaching the retina can be represented as

$$\ell_e(\mathbf{x}) = \ell_a\left(R_{f_e}^{-1}P_{d_e}^{-1}\mathbf{x}\right), \text{where } \ell_a(\mathbf{x}) = \text{rect}\left(\frac{x}{a}\right)\ell_o(\mathbf{x}),$$

and a is the diameter of the pupil. To understand the effect of the aperture, we analyze a more general situation where the light field is multiplied with a general function $h(\mathbf{x})$ and transformed by an



(a) light-field parameterization (b) at display (c) propagating d_o (d) refracted by the focus-tunable lens and creating a virtual display at v_i

(e) refracted by the lens in our eye, whose focal length is f_e (f) propagating d_e to the retina (g) aperture function propagated d_e to the retina (h) cross-correlation of (f) and (g) results in the light field at the retina (shaded blue). Our eye sees the slice along f_x.

Fig. 1. Fourier transform of the 2-dimensional light field at each stage of a multifocal display. The display is assumed to be isotropic and has pixels of pitch Δx. (a) Each light ray in the light field is characterized by its intercepts with two parallel axes, x and u, which are separated by 1 unit, and the origin of the u-axis is relative to each individual value of x. (b) With no angular resolution, the light field spectrum emitted by the display is a flat line on f_x. We focus only on the central part ($|f_x| \leq \frac{1}{2\Delta x}$). (c) The light field propagates d_o to the tunable lens, causing the spectrum to shear along f_u. (d) Refraction due to the lens corresponds to shearing along f_x, forming a line segment of slope $-v_i$, where v_i is the depth of the focal plane. (e,f) Refraction by the lens in our eye and propagation d_e to the retina without considering the finite aperture of the pupil. (g) The spectrum of the pupil function propagates d_e to the retina. (h) The light field spectrum on the retina with a finite aperture is the 2-dimensional cross-correlation between (f) and (g). According to Fourier slice theorem, the spectrum of the perceived image is the slice along f_x, shown as the red line. The diameter of the pupil and the slope of (f), which is determined by the focus of the eye and the virtual depth v_i, determine the spatial bandwidth, W, of the perceived image.

invertible T with unit determinant. By multiplication theorem, we have

$$\ell_a(\mathbf{x}) = h(\mathbf{x}) \times \ell_o(\mathbf{x}) \overset{\mathcal{F}}{\longleftrightarrow} L_a(\mathbf{f}) = H(\mathbf{f}) * L_o(\mathbf{f}).$$

Thereby,

$$L_a(T\mathbf{f}) = \int L_o(\mathbf{p}) H(\mathbf{p} - T\mathbf{f})\, d\mathbf{p} = \int L_o(\mathbf{p}) H\left(T\left(T^{-1}\mathbf{p} - \mathbf{f}\right)\right) d\mathbf{p}$$
$$= \int L_o\left(T(\mathbf{q}+\mathbf{f})\right) H(T\mathbf{q}) \left|\frac{\partial \mathbf{p}}{\partial \mathbf{q}}\right| d\mathbf{q} = L_o^{(T)} \otimes H^{(T)}(\mathbf{f}), \quad (3)$$

where we use a change of variable by setting $\mathbf{q} = T^{-1}\mathbf{p} - \mathbf{f}$, and the last equation holds because $\left|\frac{\partial \mathbf{p}}{\partial \mathbf{q}}\right| = \det T = 1$. Equation (3) relates the effect of the aperture directly to the output light field at the retina: The spectrum of the output light field is the cross correlation between the transformed (refracted and propagated) input spectrum with full aperture and the transformed spectrum of the aperture function. The result is important since it significantly simplifies our analysis, and as a result, we are able to derive an analytical expression of spatial resolution and number of focal planes needed.

In our scenario, we have $T = \left(R_{f_e}^{-1} P_{d_e}^{-1}\right)^{-\top}$. For a virtual display at v_i, $\ell_o(\mathbf{x})$ is a line segment of slope $-v_i$ within $x \in [\frac{-1}{2\Delta x_i}, \frac{1}{2\Delta x_i}]$, where $\Delta x_i = \left|\frac{v_i}{d}\right| \Delta x$ is the magnified pixel pitch. According to Equation (3), $L_e(\mathbf{f}) = L_a(T\mathbf{f})$ is simply the cross correlation of $L_o(T\mathbf{f})$ and $\mathrm{sinc}(T\mathbf{f})$. After transformation, $L_a(T\mathbf{f})$ is a line segment of slope $\frac{d_e v_i - (d_e + v_i) f_e}{v_i - f_e}$, where $|x| \leq \left|\left(\frac{v_i}{f_e} - 1\right) \frac{1}{\Delta x_i}\right|$. Similarly, $\mathrm{sinc}(T\mathbf{f})$ is a line segment with slope $-d_e$ within $|x| \leq \frac{1}{2a}$. Note that we only consider $|x| \leq \frac{1}{2a}$ because the cross-correlation result at the boundary has value $\mathrm{sinc}(0.5) \times \mathrm{sinc}(0.5) \approx 0.4$. Since $\mathrm{sinc}(x)$ function is monotonically decreasing for $|x| \leq 1$, the half-maximum spectral bandwidth ($|L_e(\mathbf{f})| = 0.5$) must be within the region. Let the depth the eye is focusing at be v. We have $\frac{1}{v} + \frac{1}{d_e} = \frac{1}{f_e}$. When $v = v_i$, we can see from the above expression that $L_a(T\mathbf{f})$ is a flat segment within $|f_x| \leq \frac{1}{2M\Delta x}$, where $M = \frac{d_e}{d_o}$ is the overall magnification caused by the focus-tunable lens and the lens of the eye. From Fourier slice theorem, we know that the spectrum of the image is simply the slice $L_a(T\mathbf{f})$ along f_x. In this case, the aperture has no effect to the final image, since the cross correlation does not extend

or reduce the spectrum along f_x, and the final image has the highest spatial resolution $\frac{1}{2M\Delta x}$.

Suppose the eye does not focus on the virtual display, or $v \neq v_i$. In the case of a full aperture ($a \to \infty$), the resulted image will be a constant DC term (completely blurred) because the slice along f_x is a delta function at $f_x = 0$. In the case of finite aperture diameter a, with a simple geometric derivation (see Figure 1h), we can show by simple geometry that the bandwidth of the f_x-slice of $L_e(\mathbf{f})$, or equivalently, the region $\{f_x | L_e(f_x, 0) \geq 0.5\}$, is bounded by $|f_x| \leq W$. And we have

$$W = \begin{cases} \frac{d_o}{2\Delta x d_e}, & \text{if} \left|\frac{1}{v_i} - \frac{1}{v}\right| \leq \frac{\Delta x}{a d_o} \\ \frac{1}{2ad_e}\left|\frac{1}{v} - \frac{1}{v_i}\right|^{-1}, & \text{otherwise.} \end{cases} \tag{4}$$

Thereby, based on Fourier slice theorem, the bandwidth of the retinal images is bounded by W.

B OTHER DISCUSSIONS

Color. Color display can be implemented by using a three color LED and cycling through them using time division multiplexing. This would lead to loss in time-resolution or focal stack resolution by a factor of 3. This loss in resolution can be completely avoided with OLED-based high speed displays since each group of pixels automatically generate the desired image at each focal stack.

Stereo virtual display. The proposed method can be extended to support stereo virtual reality displays. The most straight-forward method is to use two sets of the prototypes, one for each eye. Since all focal planes are shown in each frame, there is no need to synchronize the two focus-tunable lenses. It is also possible to create a stereo display with a single focus tunable lens and a single tracking module; the design for this is shown in Figure 2. This design trades half of the focal planes to support stereo, and thereby, only requires one set of the prototype and additional optics. Polarization is used to ensure that each eye only sees the scene that is meant to see.

C SIMULATED SCENE

Figure 3 shows the simulated images of Figure 11 in the paper with full field-of-view. There are 28 resolution charts located at various depths from 0 to 4 diopters (as indicated by beneath each of them). In the figure, we plot the ground-truth rendered images and simulated retinal images when focused on 0.02 diopters and 0.9 diopters. Rest of the focus stack can be seen in the supplemental video.

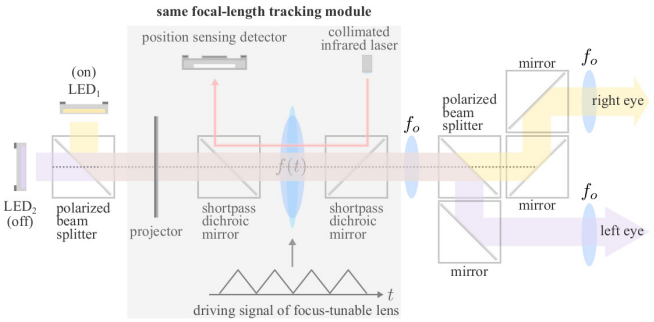
REFERENCES

Eugene Hecht. 2002. *Optics.* Addison-Wesley.

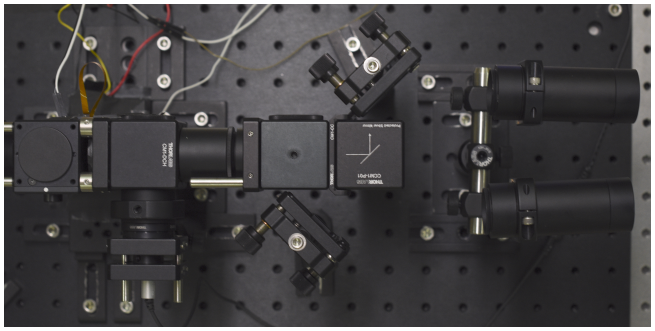
Olivier Mercier, Yusufu Sulai, Kevin Mackenzie, Marina Zannoli, James Hillis, Derek Nowrouzezahrai, and Douglas Lanman. 2017. Fast Gaze-contingent Optimal Decompositions for Multifocal Displays. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 237:1–237:15.

Rahul Narain, Rachel A Albert, Abdullah Bulbul, Gregory J Ward, Martin S Banks, and James F O'Brien. 2015. Optimal Presentation of Imagery with Focus Cues on Multi-plane Displays. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 59:1–59:12.

Qi Sun, Fu-Chung Huang, Joohwan Kim, Li-Yi Wei, David Luebke, and Arie Kaufman. 2017. Perceptually-guided Foveation for Light Field Displays. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 192:1–192:13.

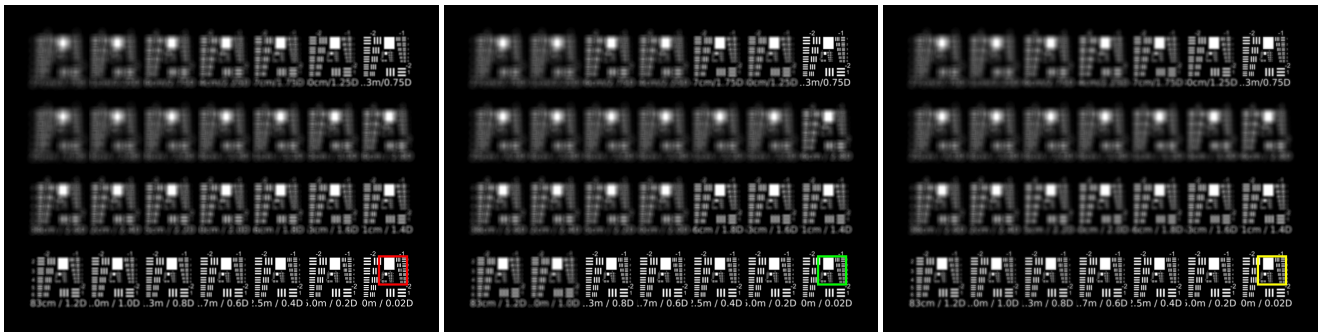


(a) Schematic

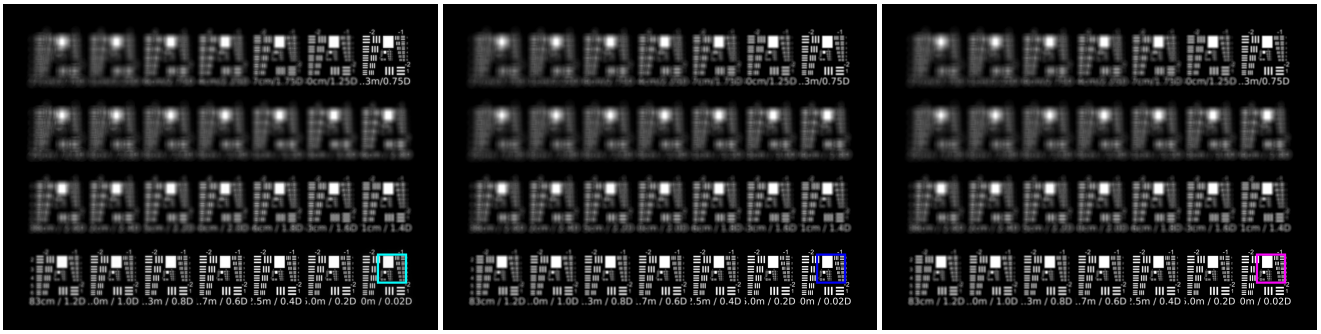


(b) The front portion of the stereo version of the prototype

Fig. 2. Supporting stereo with a single focus-tunable lens and focus-length tracking module. The design utilizes two LEDs as light sources of the DMD projector. Two polarized beam splitters are used to create dedicated light path for LED_1 (to the right eye) and LED_2 (to the left eye). To show the content on the DMD to the right eye, only LED_1 is turned on, and vice versa. To account for the extra distance created by the optics, we use two 4f systems (sharing the first lens) with $f = 75$ mm to bring both eyes virtually to the aperture of the focus-tunable lens.

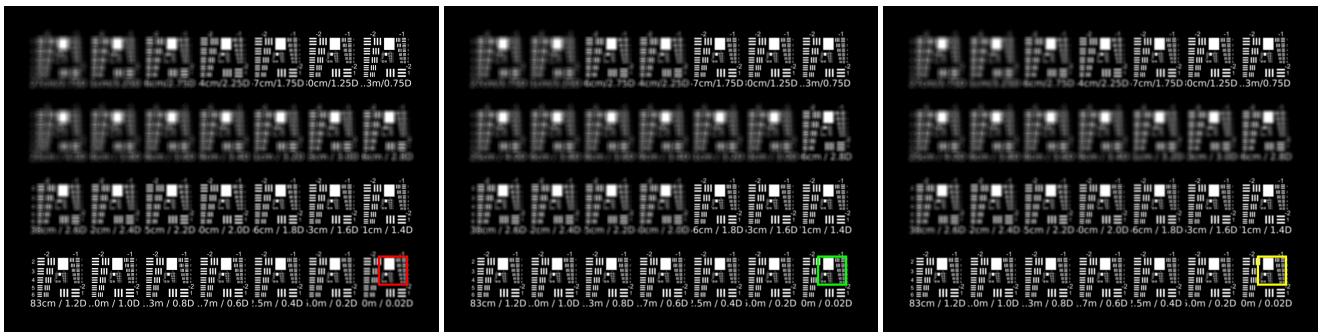


ground truth | 4-plane, direct quantization | 4-plane, linear filtering

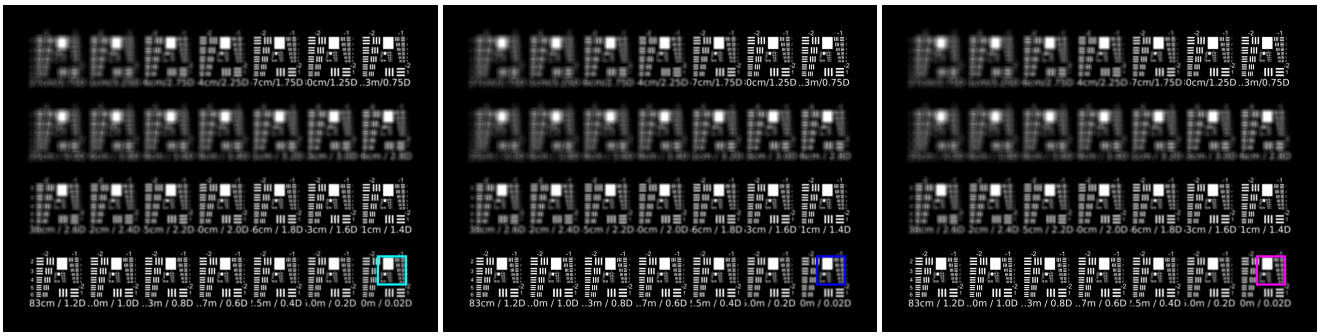


4-plane, optimization-based filtering | 40-plane, direct quantization | 40-plane, optimization-based filtering

(a) Camera focuses at 0.02 diopters (off-plane location of the 4-plane display)



ground truth | 4-plane, direct quantization | 4-plane, linear filtering



4-plane, optimization-based filtering | 40-plane, direct quantization | 40-plane, optimization-based filtering

(b) Camera focuses at 0.9 diopters (off-plane location for both displays)

Fig. 3. Simulation results of 4-plane and 40-plane multifocal displays with direct quantization, linear depth filtering, and optimization-based filtering. The indicated regions are used to plot Figure 11 in the paper.