nature portfolio

https://doi.org/10.1038/s41586-025-09752-y

Supplementary information

Glasses-free 3D display with ultrawide viewing range using deep learning

In the format provided by the authors and unedited

Supplementary Information for

Glasses-Free 3D Display with Ultrawide Viewing Range using Deep Learning

Weijie Ma, Zhangrui Zhao, Canyu Zhao, Wanli Ouyang, Han-Sen Zhong

This PDF file includes

Contents

List of Figures

List of Tables

Supplementary Methods

1 SBP Utilization Analysis with Display Outcomes

1.1 Compromise on static passive utilization of limited SBP

Since the inception, 3D displays have been stuck adapting to the inherent scarcity of space-bandwidth product (SBP), leading to various display architectures that represent passive display outcomes in the trade-off between wide viewing angles and large sizes (Supplementary Table 1). Holographic displays are a typical technology for achieving wide-angle 3D visualization in extremely small scales. Due to the significant compression of display space, they can present images with high spatial frequency. With the advancement of updatable holography [1–3], computer-generated holograms with high fidelity have been achieved using spatial light modulators (SLMs). These holograms with optical information completeness allow for wide-angle 3D observation. However, these displays typically range in size ($\sim$1-2 cm^2), not suitable for normal human viewing and often requires close, monocular viewing [4].

In parallel, for more practical display purposes, many studies have explored achieving 3D displays at common screen sizes, but SBP limitations make reproducing the entire light field at such scales unfeasible. They rely on binocular parallax, where the brain generates stereoscopic perception from different light paths received by each eye [5], referred to as automultiscopic display that enables limited effectual views due to SBP scarcity. Thus, two solutions have emerged into view-segmented and view-dense choices. The former focuses on practical implementation of a broadish viewing range by fragmenting the imaging space into separate and reused directions, using optical-path alterations with a tailored flat panel [6, 7], such as lenticular lenses [8, 9], parallax barrier [10, 11], multiple projectors [12, 13] and eye-tracking integration [14]. The trade-off is introducing viewpoint discontinuity, leading to imaging jitter and sensing mismatch under ocular natural motion due to discrete view transitions [15, 16], resulting in visual distortion, broken immersion [17] by interpupillary distance (IPD) mismatch, and even visual-vestibular conflict that causes motion sickness [18]. The latter prioritizes the viewing smoothness and realism by stacking screens into a compressive light field with direct depth cues [19–21], maintaining parallax modulation with omnidirectional pixel-level continuity [22]. Yet, this densification comes with the cost of a highly restricted viewing angle and the incapability of closer-range views by treating binocular projections as simple shifts.

As a result, these SBP passive utilization in existing display outcomes presents significant challenges for desirable glasses-free 3D emergence, which requires seamless imaging transitions across full parallaxes. Beyond basic planar types (horizontal and vertical), full parallaxes encompass advanced attributes like motion, radial, and focal components, representing movement support, convergence, and accommodation, respectively [26]. Besides, seamless transitions necessitate motion parallax to pixel-level variation continuity [27]. Due to SBP limitations, current architectures fail to meet all parallax requirements. Updatable holography struggles to support large movement and radial parallax due to its extreme scale limitations. View-segmented

	Paradigm	Holographic displays	Automultiscopic displays		EyeReal
			View-segmented	View-dense	
Scalability	Large display size	×	✓	✓	✓
	Seamless wide angle	✓	×	×	✓
	Variable viewing distance	×	×	×	✓
	Variable movement range	×	×	×	✓
Parallax	Horizontal parallax	✓	✓	✓	✓
	Vertical parallax	✓	×	✓	✓
	Radial parallax	✓	×	✓	✓
	Motion parallax	✓ (Restricted)	✓ (Discrete)	×	✓
	Focal parallax	✓	×	✓	✓
Cost	Computational efficiency	Realtime (monocular)	Preset (specific positions)		Realtime
	Hardware cost	High	Moderate or high	Low	Low

Supplementary Table 1 Comprehensive comparison of different modern 3D display paradigms. "✓" denotes support; "×" denotes lack of support.

	Method	Neural CGH Synthesis (Shi et al. [2])	Neural Light-field Synthesis (Maruyama et al. [23]; Sun et al. [24, 25])	EyeReal
Scalability	Large display scale	×	✓	✓
	Large viewing angle	✓	×	✓
	Variable viewing distance	×	×	✓
	Variable movement range	×	×	✓
Latency		Low	Low	Low
Image fidelity		High (monocular)	High (parallel eye)	High

Supplementary Table 2 Comparison of different neural 3D display representatives and EyeReal. "✓" denotes support; "×" denotes lack of support.

automultiscopy has to sacrifice some spatial parallax (such as vertical or radial) due to unnatural, discrete parallax formation, and its single flat-panel nature loses focal parallax, leading to potential vergence-accommodation conflict (VAC) with fatigue or dizziness [28–30]. Symmetrically, view-dense choices, with their narrow fixed viewing regions and non-negligible rotational deformation, also cannot support obvious movement and radial parallax. All the features above still persist in current displays, and subsequent progresses tend to offer incremental improvements, even including AI-driven neural methods (Supplementary Table 2), which optimize latency and display quality within the confines of existing architectures.

1.2 Related paradigm development for autostereoscopy

Desirable autostereoscopy aims to reproduce light fields at natural viewing ranges, which means achieving a wide field of view at large imaging sizes. This has remained a long-standing challenge in modern optics because it requires an extremely high SBP, which is beyond the capabilities of existing optical display devices and data transmission technologies restricted by the Lagrange invariant [31]. Extensive research in the near-eye range [2, 32–37] has been studied to display realistic images with remarkable performance. With the advancement of holography, computer-generated holograms with exceptional fidelity have been achieved using SLMs or analogous optics [33]. These holograms with completeness of optical information enable wide-angle 3D content observation. However, fundamental SBP limitations confine the achievable imaging size to the centimeter scale ($\sim$1-2 cm^2), necessitating impractically close viewing distances that deviate from natural viewing experience [4]. Even with state-of-the-art SLMs, regardless of cost, scaling to commonly-used display dimensions (*e.g.*, 15-27 inches, $\sim$0.1-0.2 m^2) remains physically impossible due to astronomical SBP requirements.

The dilemma of achieving naturally viewable glasses-free 3D display within limited SBP has been addressed by trading viewing angles for larger imaging size. Autostereoscopic multi-view display, referred to as automultiscopic display, which enables glasses-free 3D viewing by directing different images to each eye, offers a promising approach toward this goal [27]. It relies on binocular parallax, where the human brain generates stereoscopic perception when the eyes receive light rays from different paths [5]. Discrete-parallax light-field approximations fragment the imaging space into separate directions based on parallax barriers [10, 11, 38], integral imaging [15, 30, 39, 40], multiple projectors [12, 41], visual aberration correction [42, 43], directional backlighting [44, 45], lenticular lenses [8, 9] and different variants of their combinations [7, 46, 47]. These methods maintain relatively-wide viewing angles through optical-path alteration using the tailored flat panel, but inevitably introduce viewpoint discontinuity. This forced discretization can cause apparent imaging jitter or graininess under viewer motion due to discrete view transitions [15, 16] with IPD mismatch, leading to visual distortion and broken immersion [17]. Additionally, the single flat-panel nature and unnatural parallax formation usually lead to VAC [28–30] due to the mismatch between the depth cues perceived by human brains and the actual fixed distance of the display. This phenomenon can potentially induce fatigue and dizziness, even in the most advanced commercial products currently available.

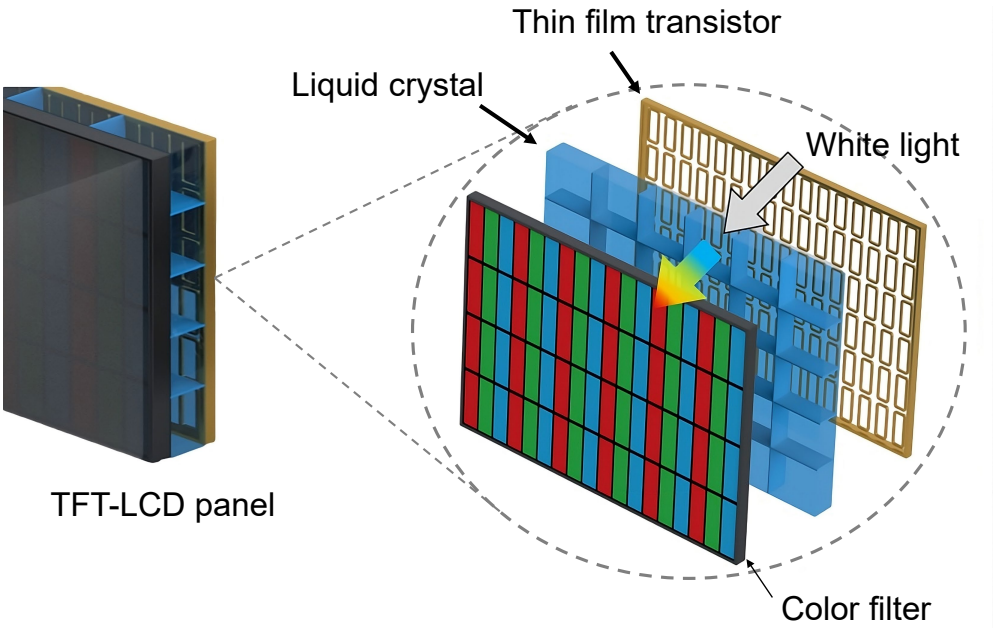
In contrast, multilayer liquid crystal displays (LCDs) with polarizers [19, 21, 48–50] model direct depth cues by stacking screens into a compressive light field, avoiding IPD mismatch and VAC [22, 51]. Following Malus' law, they achieve this through linear aggregation of luminous intensity [20, 21] or optical phase [48, 52, 53] at hierarchical depths, enabling omnidirectional pixel-level modulation and continuous stereo parallax. However, due to SBP constraints, this densification results in a highly restricted viewing angle range. Techniques like tensor displays [54, 55], focus guidance [56, 57], and tailored weighting [25, 58] have been developed to optimize stereo perception within these limited angles, but they rely on over-idealistic long-distance viewing assumptions (approximating parallel eye) where binocular projections are simplified as narrow (pixel-level) shifts. This leads to incompatibility at close proximity with nonnegligible rotational deformation and making radial stereopsis over large ranges unattainable. Furthermore, the widely-used iterative optimization methods [11, 54, 59, 60] based on nonnegative matrix and tensor factorization (NMF and NTF) [61] for computing optimal optical patterns are computationally intensive, resulting in slow display speeds that hinder variable movement within the viewing space. Despite over five decades of research, all existing approaches aforementioned face fundamental trade-offs under SBP constraints, limiting the simultaneous achievement of large-scale imaging and wide-angle viewing required by truly glasses-free 3D displays.

2 Light-Field Display with Multi-Layer LCDs

2.1 Thin-film transistor liquid crystal display

In our system, we utilize Thin-Film Transistor Liquid Crystal Displays (TFT-LCDs), a sophisticated evolution of traditional LCD technology. TFT-LCDs are distinguished by their incorporation of thin-film transistors (TFTs), which enhance the display's performance by enabling precise control over individual pixels. The architecture of a TFT-LCD is composed of several critical components (Supplementary Figure 1), including two glass substrates, a liquid crystal layer, a matrix of TFTs, pixel electrodes, color filters, common electrodes, polarizers, and a backlight system.

The two glass substrates serve as the foundational layers of the display, enclosing the liquid crystal layer. On one of these substrates, a matrix of TFTs and pixel electrodes is meticulously arranged. These TFTs function as switches, allowing for the application of exact voltages to the liquid crystal molecules within each pixel. This voltage control is crucial because it determines the orientation of the liquid crystal molecules, thereby modulating the amount of light that passes through each pixel and ultimately creating the desired image on the screen. On the opposite substrate, color filters and common electrodes are positioned. The color filters are configured to create red, green, and blue sub-pixels, which are the fundamental building blocks of the full color spectrum displayed on the screen. By combining these sub-pixels in varying intensities, the TFT-LCD can produce a wide range of colors, rendering images with high fidelity and vibrant hues. To manage the polarization of light as it enters and exits the display, polarizers are affixed to the outer surfaces of the glass substrates. These polarizers are essential for ensuring that the light passing through the liquid crystal layer is appropriately aligned, contributing to the display's contrast and clarity.



Supplementary Figure 1 Structure of a thin-film transistor liquid crystal display (TFT-LCD). This diagram illustrates the key components of a TFT-LCD, including the glass substrates, liquid crystal layer, TFT matrix, color filters, and backlight system. The arrows represent the path of light as it passes through the display white light enters from the backlight, which is then modulated by the liquid crystal layer based on voltage control from the TFTs. As the light passes through the color filters, it is split into red, green, and blue components, which combine to form the full-color image displayed on the screen.

2.2 Light field display based on Malus' law

In TFT-LCD's liquid crystal layer, the orientation of the liquid crystal molecules is controlled by the electric fields generated by the TFTs. By varying the applied voltage, the alignment of the liquid crystal molecules is adjusted, altering the polarization state of the light and thus the intensity of the light that exits the display. The operation of TFT-LCDs is deeply intertwined with Malus's law, which governs the intensity of polarized light passing through a LCD. According to Malus's law, the intensity of the transmitted light I is a function of the incident light intensity I_0 and the angle θ between the polarization direction of the light and the transmission axis of liquid crystal, which can be formulated by

$$I = I_0 \sin^2(\theta) \tag{1}$$

Here we throw a brief proof for the above equation based on Jones calculation [62]. More details can be referred in some reviews [48, 63–65]. A single LCD panel can be modeled as a polarization rotator, which is approximated by applying a linear rotation to the incident polarization state. The Jones matrix $R(\theta)$, representing a rotation by θ, is expressed as

$$R(\theta) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \tag{2}$$

Additionally, the Jones matrix $J(\xi)$, which corresponds to a counterclockwise rotation of an optical element by an angle ξ, is given by the transformation $R(-\xi)JR(\xi)$, where J is the Jones matrix of the unrotated element. Utilizing these principles, we derive the expression for the normalized intensity $I_{R1}(\phi, \xi)$ for a single polarization rotator enclosed by two linear polarizers, as a function of the polarization state rotation angle ϕ and the angle ξ between the axes of the front and rear polarizers

$$I_{R1}(\phi, \xi) = I_0 \left\| \left(R(-\xi) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} R(\xi) \right) R(-\phi) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|^2 = I_0 \cos^2(\phi - \xi) \tag{3}$$

Here the vector norm is l_2 norm. In this context, the rear polarizer, closest to the backlight, is assumed to be horizontally aligned, while the front polarizer is allowed to rotate freely. By convention, we assume the polarization rotator induces a counterclockwise rotation. For crossed polarizers such as a horizontal polarizer and a vertical polarizer (i.e., $\xi = \pi/2$), Equation 3 simplifies to

$$I_{R1}\left(\phi, \frac{\pi}{2}\right) = I_0 \sin^2(\phi) \tag{4}$$

thereby confirming Malus' law as illustrated in Equation 1.

Then we apply the Jones calculus to analyze multilayer LCDs as multilayer polarization rotators. The Jones matrix $J_{RN}(\phi_1, \phi_2, \cdots, \phi_N)$, representing a composition

of N polarization rotators, is given by

$$\begin{aligned} J_{RN}(\phi_1, \phi_2, \cdots, \phi_N) &= R(-\phi_N)R(-\phi_{N-1})\cdots R(-\phi_1) \\ &= \begin{pmatrix} \cos(\sum_{n=1}^{N}\phi_n & -\sin\sum_{n=1}^{N}\phi_n \\ \sin(\sum_{n=1}^{N}\phi_n & \cos\sum_{n=1}^{N}\phi_n \end{pmatrix} \end{aligned} \tag{5}$$

where $\{\phi_1, \phi_2, \cdots, \phi_N\}$ represent the incremental polarization state rotations induced at each layer. A comparison of Equation 5 with Equation 2 demonstrates that a sequence of polarization rotators effectively applies a counterclockwise rotation to the incident polarization state by an angle

$$\theta_{seq} = \sum_{n=1}^{N}\phi_n \tag{6}$$

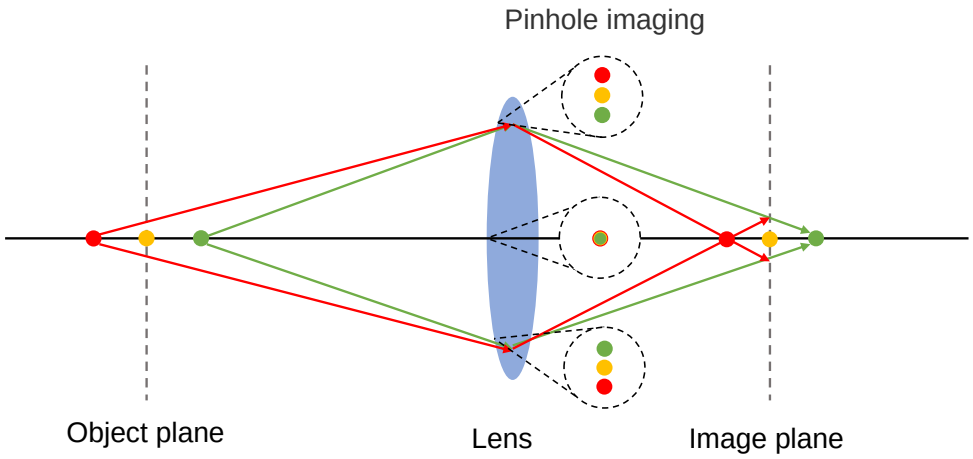
Consequently, the normalized intensity $I_{RN}(\phi_1, \phi_2, \cdots, \phi_N)$ for a N-layer polarization rotator enclosed by crossed linear polarizers is expressed as

$$I_{RN}(\phi_1, \phi_2, \cdots, \phi_N) = I_0 \left\| (0\ 1) \left(\prod_{n=1}^{N} R(-\phi_{N-n+1}) \right) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|^2 = I_0 \sin^2 \left(\sum_{n=1}^{N}\phi_n \right) \tag{7}$$

2.3 Focal parallax formulation in multi-layer LCDs

Focal parallax is a depth-related visual phenomenon that occurs when the eye or camera focuses on an object at a specific depth, causing objects at other depths to become blurred. This effect results from the way light rays from different depths interact with the imaging system. When the eye or lens focuses on one point, light rays from that point converge sharply onto the image plane or retina, rendering the point in sharp focus. However, light rays from objects at different depths do not converge in the same way, leading to their appearance as defocused or blurred, often seen as larger, soft spots. This is due to the varying angular divergences of light from these objects, which cause differential projections on the image plane. In optical systems, such as cameras or the human eye, this effect is more pronounced when viewed through a small aperture, where the focal point becomes distinct while points at other depths spread out as blur.

To enable the camera or eyes to freely select the focus position, it is essential that the light emitted from points at different depths in the displayed content exhibits varying degrees of divergence. In the context of light field displays, this necessitates the presence of continuous parallax in the light field entering the camera or human eye, meaning different perspectives at different positions across the lens, as shown in Supplementary Figure 2. This capability is unattainable with view-segmented flat-panel displays, as they can only observe the image corresponding to the central view within a spatial domain. In contrast, our system is stacked by layered LCDs, allowing for focal parallax, a capability beyond the reach of view-segmented flat-panel displays.



Supplementary Figure 2 A schematic diagram illustrates the imaging process with focal parallax. When the focus point overlaps with the yellow point, the green and red points become defocused, spreading into larger spots on the camera sensor. This phenomenon arises due to points with distinct depth having different perspective relationships. This effect can be observed when imaging these three points through a small aperture within the optical system.

Besides, unlike traditional displays, where all content appears on a single plane and lacks depth cues, our light field system allows the viewer's eyes to freely shift focus across different depths. This is achieved by emitting light rays that continuously vary in perspective across viewpoints, simulating the way real-world light behaves. Traditional view-segmented flat-panel systems, by contrast, are limited in this regard, offering only a single focal depth and failing to provide true focal parallax, which can result in visual artifacts like vergence-accommodation conflict (VAC).

2.4 Depth of field analysis for VAC avoidance

The capability of focal parallax to eliminate VAC with perceptual tolerance can be derived by Fourier optical analysis, as detailed below.

We define the maximum spatial detail that can be rendered on a given depth plane as the depth-dependent spatial cutoff frequency, denoted by $\omega_{\xi_{max}}(d_o)$, where d_o represents the distance between that plane and the main display plane. Through frequency domain analysis, we first derive the light field spectrum for an N-layer screen

$$\tilde{l}(\omega_x, \omega_v) = \frac{1}{M} \sum_{m=1}^{M} \tilde{b}_m(\omega_x, \omega_v) * \left[\underset{n=1}{\overset{N}{*}} \tilde{f}_m^{(n)}(\omega_x) \delta(\omega_v - \frac{d_n}{d_r}\omega_x) \right], \tag{8}$$

where ω_x and ω_v represent spatial frequency and angular frequency, respectively, the symbol $*$ denotes the convolution operation, and $\tilde{b}_m$ and $\tilde{f}_m^{(n)}$ are the spectra of the backlight light field for the m-th frame and the n-th layer's transmittance function, respectively. δ is the Dirac delta function.

To determine the spatial cutoff frequency at a specific depth d_0, we employ the spectral intersection principle. According to light field theory, the spectral energy of all light rays generated by an ideal diffuse surface at depth d_0 is concentrated along a straight line, defined by the equation $\omega_v = (d_0/d_r)\omega_x$. Therefore, for a display system to reproduce details on this plane without distortion, its own light field spectral support must completely cover this line. The system's performance bottleneck, specifically the spatial cutoff frequency $\omega_{\xi_{max}}(d_o)$, is determined by the intersection point of this depth characteristic line with the boundaries of the display's spectral support. For an N-layer screen, the total spectral support forms a 2N-sided convex polygon. By intersecting this spectral region with the line $\omega_v = \frac{d_o}{d_r}\omega_x$, the maximum spatial frequency $\omega_{\xi_{max}}(d_o)$ can be determined.

The spatial Nyquist frequency of a single display layer $\omega_0 = \frac{1}{2p}$, determined by the pixel pitch p, represents the maximum spatial frequency that can be faithfully represented without aliasing on each layer, and thus sets the angular and spatial resolution limits in the frequency domain for the multilayer display system. When the observation plane is in the near field, i.e., $|d_o| \leq T = (N-1)\Delta d$, the intersection line falls within the slanted edge segment. When $|d_o| > T$, the intersection line instead intersects the vertical edges of the spectral envelope. Therefore, $\omega_{\xi_{max}}(d_o)$ is formulated as

$$\omega_{\xi_{max}}(d_o) = \begin{cases} \frac{N\Delta d}{\Delta d + |d_o|}\omega_0, & |d_o| \leq (N-1)\Delta d \\ \frac{(N-1)\Delta d}{|d_o|}\omega_0, & |d_o| > (N-1)\Delta d \end{cases} \tag{9}$$

To ensure that the display presents content at its maximum spatial fidelity, it is necessary that the cutoff frequency satisfies $\omega_{\xi_{max}}(d_o) \geq \omega_0$. Notably, in the near-field region, the cutoff frequency exceeds ω_0. Therefore, to fully preserve the finest details of the content, the object distance d_o needs to satisfy $|d_o| \leq (N-1)\Delta d$.

According to the accommodation model, a visually comfortable display for the human eye requires an accommodation range of at least $\Delta D \geq 0.3D$ [28]. The dioptric difference between a virtual object at depth $L + |d_o|$ and the screen plane (at distance L from the eye) is

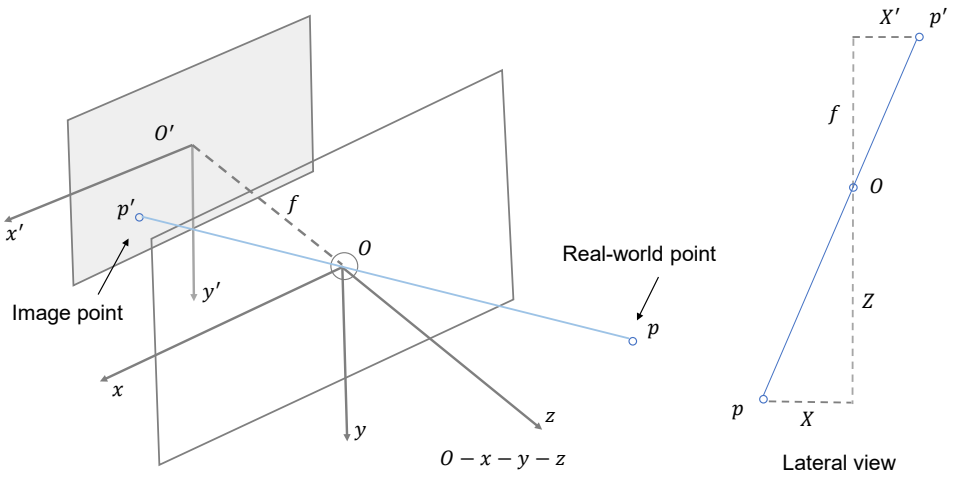
$$\Delta D = \left| \frac{1}{L + |d_o|} - \frac{1}{L} \right| \tag{10}$$

Our original configuration, a three-layer screen with an inter-layer spacing of 3 cm and a viewing distance $L = 0.95$ m, provides only approximately 0.183 diopters of accommodation disparity. This falls short of the 0.3 threshold necessary to meet the recommended accommodation margin. However, we have investigated more recent hardware implementations. For example, a six-layer display prototype with 4 cm inter-layer spacing and 1920×1080 resolution has been successfully demonstrated [66]. Although the screen size is not explicitly reported in that study, using our derived general expression for the cutoff frequency, we estimate that such a 6-layer configuration (with 4 cm spacing) would require a viewing distance of approximately 0.72 m to support a retinal-resolution accommodation range of ±0.3 D. This corresponds to a screen diagonal of roughly 18 inches. These parameters are well within the range of typical and practically realizable display form factors, suggesting that our method is fundamentally capable of meeting the recommended standards for VAC comfort.

3 Eye-Lightfield Standardization

3.1 Pinhole camera model and its imaging geometry

We adopt the pinhole camera model as the mathematical approximation of eye imaging. In the pinhole camera model, the imaging process can be described mathematically using a simple coordinate system. Assume that the pinhole is located at the origin of the coordinate system, with the image plane located at a distance behind the pinhole along the z-axis, where is the focal length, analogous to the distance from the pinhole to the imaging surface. To develop a comprehensive mathematical model for the process of projecting a three-dimensional world scene onto a two-dimensional image, we begin by adopting the camera's coordinate system as the reference frame. This choice allows us to establish the mathematical relationships governing the projection process with precision. The camera-centered reference system is defined as a right-handed coordinate system, where the origin of the coordinates is positioned at the camera's optical center, commonly referred to as the pinhole in the pinhole camera model. In this coordinate system, the z-axis corresponds to the camera's central optical axis, which is a straight line extending from the optical center and perpendicular to the image plane. In this framework, the image plane is mathematically described by the equation $Z = f$, where f represents the focal length of the camera. As depicted in the left portion of Supplementary Figure 3, we denote the camera coordinate system



Supplementary Figure 3 The geometrical imaging process of the pinhole camera model. The left panel depicts the 3D camera-centered coordinate system $O - x - y - z$, where the pinhole (optical center) is at O, and the image plane is located at $Z = f$ (focal length). A real-world point P is projected onto the image plane as p' via rays passing through O. The dashed lines show the projection geometry. The right panel provides a lateral view, showing the proportional relationship among p, p' and the focal plane through similar triangles.

by $O-x-y-z$, where O indicates the optical center. Let $[X, Y, Z]^T$ represent the coordinates of a point P in the real world, and $[X', Y', Z']^T$ denote the coordinates of the corresponding image point P'. By applying the principle of similar triangles, illustrated in the right portion of Supplementary Figure 3, we derive the following relationship

$$\frac{Z}{f} = -\frac{X}{X'} = -\frac{Y}{Y'} \tag{11}$$

Here, the negative sign indicates the direction of the coordinate axis, signifying that the image appears inverted relative to the object. To simplify the analysis and eliminate the negative sign, we relocate the imaging plane from behind the camera to the front. With this adjustment, the relationship becomes

$$\frac{Z}{f} = \frac{X}{X'} = \frac{Y}{Y'} \tag{12}$$

This leads to the solution for the coordinates of P', given by

$$\begin{cases} X' = f\dfrac{X}{Z} \\ Y' = f\dfrac{Y}{Z} \end{cases} \tag{13}$$

Next, we map the calculated coordinates on the image plane to the pixel coordinate system. The pixel plane, denoted by $O-u-v$, is fixed on the physical imaging plane. Assume that the pixel-level coordinates of P' are $[u, v]^T$. We further scale the coordinates by factors of α along the u-axis and β along the v-axis. Additionally, the origin of the pixel coordinate system is shifted by $[c_x, c_y]^T$. The relationship between the coordinates of P' and the pixel coordinates is then expressed as

$$\begin{aligned} u &= \alpha X' + c_x \\ v &= \beta Y' + c_y \end{aligned} \tag{14}$$

By substituting the relationship between P and P' derived earlier, i.e.,

$$\begin{aligned} u &= \alpha f\frac{X}{Z} + c_x = f_x\frac{X}{Z} + c_x \\ v &= \beta f\frac{Y}{Z} + c_x = f_y\frac{Y}{Z} + c_y \end{aligned} \tag{15}$$

Here, we replace αf and βf with f_x and f_y respectively, where f_x and f_y represent the focal lengths in pixels along the u and v axes. Using homogeneous coordinates, we can write the above equations in matrix form

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{Z}\begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{1}{Z}\mathbf{KP} \tag{16}$$

Z can also be written to the left of the above equation, which is transformed into

$$Z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{KP} \tag{17}$$

In this equation, $\mathbf{K}$ represents the camera's intrinsic matrix, encapsulating the internal parameters of the camera.

To represent the transformation from any coordinate system to the image pixel coordinate system, we first transform the point P_ω in the world coordinate system to the corresponding point P in the camera coordinate system. This transformation is achieved through the camera pose, described by the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$. These two components can be combined into a single transformation matrix $\mathbf{T} = [\mathbf{R}|\mathbf{t}]$, which belongs to the Special Euclidean group $SE(3)$. The $SE(3)$ group is defined as follows

$$SE(3) = \left\{ \mathbf{A} \;\middle|\; \mathbf{A} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}, \mathbf{R} \in \mathbb{R}^{3\times 3}, \mathbf{t} \in \mathbb{R}^3, \mathbf{R}^T\mathbf{R} = \mathbf{R}\mathbf{R}^T = \mathbf{I}, |\mathbf{R}| = 1 \right\}$$

Here, $\mathbf{R}$ is a 3×3 orthogonal matrix representing the rotation of the camera, and $\mathbf{t}$ is a 3×1 vector representing the translation. The properties $\mathbf{R}^T\mathbf{R} = \mathbf{I}$ and $|\mathbf{R}| = 1$ ensure that the rotation matrix is valid, preserving the lengths and angles during transformation. The relationship between the point P_ω in the world coordinate system and the point P in the camera coordinate system can be expressed as

$$\mathbf{P} = \mathbf{R}\mathbf{P}_\omega + \mathbf{t} \quad \Leftrightarrow \quad \mathbf{P}|_{homo} = \mathbf{T}\mathbf{P}_\omega|_{homo} \tag{18}$$

In this equation, P and P_ω denote the column vectors of points P and P_ω. $P|_{homo}$ and $P_\omega|_{homo}$ represent their homogeneous coordinates respectively. By substituting this relationship into the earlier derived equation for the imaging process, we obtain the complete formulation for mapping a point P_ω in the world coordinate system to its corresponding point in the pixel coordinate system

$$Z\mathbf{P}_{uv}|_{homo} = \mathbf{K}(\mathbf{R}\mathbf{P}_\omega + \mathbf{t}) \tag{19}$$

3.2 Eye coordinate system setup

The eye-camera model described in the main text is foundational for simulating the retinal imaging process using a light field coordinate system. This model provides a geometric framework to understand how the human eye perceives objects within a defined 3D space, replicating the actual position and gaze direction of the eye in the real world. The choice of a pinhole camera model stems from its simplicity and effectiveness in approximating the human eye's operation, mapping 3D points to 2D planes (retina) in an optically plausible manner. By aligning the center of the screen with the center of the light field and setting the eye's default gaze towards the origin, we create a standardized system that is easy to analyze and implement for simulating different viewing conditions. In the eye coordinate system model, the z-axis of

the camera model is defined as opposite to the gaze direction, ensuring a consistent representation of the observer's viewpoint relative to the world. This approach mirrors real-world behavior, where the visual plane extends outward from the eye into the space it is observing. The x-axis parallel to the ground ensures that objects lying on a flat surface (e.g., a table or floor) are viewed similarly by both the eye and the camera model, thereby maintaining a realistic orientation of the visual scene. Additionally, setting the y-axis perpendicular to the plane formed by the z and x axes creates a right-handed coordinate system. This ensures that the 3D transformations applied to points and vectors within this system are consistent with common geometric conventions, making mathematical operations such as rotation and translation straightforward. Below is the mathematical derivation of the projection matrix. To convert from the eye-camera coordinate system to the light field coordinate system, the transformation can be expressed as a homogeneous matrix $\mathbf{M}_e = [\mathbf{R}_e|\mathbf{t}_e]$, where $\mathbf{R}_e$ represents the rotation matrix and $\mathbf{t}_e$ represents the translation vector. This matrix maps the eye's position and orientation within the light field, enabling the simulation of retinal imaging in the light field system. The rotation matrix $\mathbf{R}_e$ is derived based on three orthogonal vectors $\mathbf{r}_x$, $\mathbf{r}_y$, and $\mathbf{r}_z$, which correspond to the x, y, and z axes of the eye camera coordinate system, respectively. The vector $\mathbf{r}_z$ is defined as the vector from the origin of the light field coordinate system to the eye position, representing the gaze direction. The vector $\mathbf{r}_x$ is perpendicular to $\mathbf{r}_z$ and lies in the plane parallel to the ground (i.e., the xOy plane of the light field coordinate system). This vector can be computed using the cross product of $\mathbf{r}_z$ and the z-axis vector of the light field system

$$\begin{cases} \mathbf{r}_x \parallel xOy \\ \overrightarrow{Oz} \perp xOy \end{cases} \implies \mathbf{r}_x \perp \overrightarrow{Oz} \tag{20}$$

Based on the above, then

$$\begin{cases} \mathbf{r}_x \perp \overrightarrow{Oz} \\ \mathbf{r}_x \perp \mathbf{r}_z \end{cases} \implies \mathbf{r}_x = \overrightarrow{Oz} \times \mathbf{r}_z \tag{21}$$

Finally, the vector $\mathbf{r}_y$ is orthogonal to both $\mathbf{r}_z$ and $\mathbf{r}_x$, ensuring the creation of an orthonormal basis for the rotation matrix:

$$\begin{cases} \mathbf{r}_x \perp \mathbf{r}_y \\ \mathbf{r}_y \perp \mathbf{r}_z \end{cases} \implies \mathbf{r}_y = \mathbf{r}_z \times \mathbf{r}_x \tag{22}$$

These vectors are normalized to ensure unit length, resulting in the rotation matrix $\mathbf{R}_e$:

$$\mathbf{R}_e = \left[\frac{\mathbf{r}_x}{\|\mathbf{r}_x\|_2}, \frac{\mathbf{r}_y}{\|\mathbf{r}_y\|_2}, \frac{\mathbf{r}_z}{\|\mathbf{r}_z\|_2} \right]^T \tag{23}$$

Here $\|\cdot\|_p$ denotes the l_p vector norm applied on these trivial vectors. This matrix describes the orientation of the eye-camera system relative to the light field, enabling accurate projection of 3D points onto the 2D retinal plane. The translation vector $\mathbf{t}_e$ represents the position of the eye in the light field coordinate system. As mentioned earlier, this vector is simply the vector OP_e, which is equivalent to $\mathbf{r}_z$, the gaze

direction

$$\mathbf{t}_e = \overrightarrow{OP_e} = \mathbf{r}_z \tag{24}$$

Thus, the projection matrix $\mathbf{M}_e = [\mathbf{R}_e|\mathbf{t}_e]$ fully captures both the position and orientation of the eye relative to the light field, allowing for the simulation of the eye's perspective on the virtual scene.

4 Additional Details of Calibration

4.1 Eye-lightfield least squares calibration

In this section, we provide a detailed explanation of the mathematical foundation and the step-by-step process involved in solving for the projection matrix $\mathbf{M}_c$ using least squares regression, as described in the main text. The goal is to minimize the alignment error between the captured camera coordinates and the corresponding light field world coordinates of calibration points, denoted as $\mathbf{c}_i$ and $\mathbf{w}_i$, respectively.

Least squares regression is a widely used optimization method for finding the best-fitting solution to an overdetermined system of linear equations. In our case, the objective is to solve for a general affine transformation matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$ and the translation vector $\mathbf{t}_c \in \mathbb{R}^3$ that define the projection matrix $\mathbf{M}_c = [\mathbf{A}|\mathbf{t}_c]$, such that the transformation from the camera coordinate system to the light field world coordinate system minimizes the discrepancy between the transformed camera coordinates and the true world coordinates of the calibration points. The matrix $\mathbf{A}$ is a general affine transformation matrix that can represent arbitrary linear transformations including rotation, scale, skew, and other geometric distortions without any constraints. Given a set of K calibration pairs $\{(\mathbf{c}_i, \mathbf{w}_i)\}_{i=1}^K$, the least squares problem can be formulated as minimizing the Euclidean distance between the transformed camera coordinates $(\mathbf{A}\mathbf{c}_i + \mathbf{t}_c)$ and the corresponding world coordinates $\mathbf{w}_i$. Mathematically, this is expressed as

$$\mathbf{A}, \mathbf{t}_c = \underset{\mathbf{A}\in\mathbb{R}^{3\times3},\, \mathbf{t}_c\in\mathbb{R}^3}{\arg\min} \sum_{i=1}^{K} \|(\mathbf{A}\mathbf{c}_i + \mathbf{t}_c) - \mathbf{w}_i\|_2^2 \tag{25}$$

where $\|\cdot\|_2$ denotes the l_2 norm, representing the Euclidean distance between the predicted and measured coordinates. This formulation allows us to quantify the error for each calibration pair and sum the squared errors across all pairs to form a single objective function. By minimizing this objective function, we can find the optimal projection matrix that minimizes the alignment error.

We model the transformation from the camera coordinate system to the light field coordinate system using a general affine transformation. The transformation is described by an unconstrained affine matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$, which can represent arbitrary linear transformations including rotation, scale, skew, and other geometric distortions, and the translation vector $\mathbf{t}_c \in \mathbb{R}^3$, which accounts for the relative displacement. For each calibration point $\mathbf{c}_i \in \mathbb{R}^3$ in the camera coordinate system, the corresponding

world coordinates $\mathbf{w}_i \in \mathbb{R}^3$ can be approximated as

$$\mathbf{w}_i \approx \mathbf{A}\mathbf{c}_i + \mathbf{t}_c \tag{26}$$

The error term for each calibration point is defined as the difference between the transformed camera coordinates and the true world coordinates, expressed as

$$\mathbf{e}_i = (\mathbf{A}\mathbf{c}_i + \mathbf{t}_c) - \mathbf{w}_i \tag{27}$$

The goal is to minimize the sum of the squared errors over all calibration points

$$\epsilon = \sum_{i=1}^{K} \|\mathbf{e}_i\|_2^2 = \sum_{i=1}^{K} \|(\mathbf{A}\mathbf{c}_i + \mathbf{t}_c) - \mathbf{w}_i\|_2^2 \tag{28}$$

To solve the least squares problem, we rewrite the transformation model in matrix form. We stack all the calibration points into matrices

$$\begin{aligned} \mathbf{C} &= \left[\mathbf{c}_1^T; \mathbf{c}_2^T; \cdots; \mathbf{c}_K^T\right] \in \mathbb{R}^{K\times 3} \\ \mathbf{W} &= \left[\mathbf{w}_1^T; \mathbf{w}_2^T; \cdots; \mathbf{w}_K^T\right] \in \mathbb{R}^{K\times 3} \end{aligned} \tag{29}$$

where each row represents a calibration point. The affine transformation model can be expressed in matrix form as

$$\epsilon = \|\mathbf{W} - (\mathbf{C}\mathbf{A}^T + \mathbf{1}_K \mathbf{t}_c^T)\|_F^2 \tag{30}$$

where $\mathbf{A}$ is a general unconstrained affine transformation matrix, $\|\cdot\|_F$ represents the Frobenius norm, and $\mathbf{1}_K$ is a $K \times 1$ column vector of ones. To solve this least squares problem directly, we can rewrite the objective function by expanding the Frobenius norm

$$\epsilon = \mathrm{tr}[(\mathbf{W} - \mathbf{C}\mathbf{A}^T - \mathbf{1}_K \mathbf{t}_c^T)^T (\mathbf{W} - \mathbf{C}\mathbf{A}^T - \mathbf{1}_K \mathbf{t}_c^T)] \tag{31}$$

To find the optimal solution, we take partial derivatives with respect to $\mathbf{A}$ and $\mathbf{t}_c$ and set them to zero. First, solving for the optimal translation vector $\mathbf{t}_c$:

$$\frac{\partial \epsilon}{\partial \mathbf{t}_c} = -2\mathbf{1}_K^T (\mathbf{W} - \mathbf{C}\mathbf{A}^T - \mathbf{1}_K \mathbf{t}_c^T) = 0 \tag{32}$$

This gives us

$$\mathbf{t}_c^T = \frac{1}{K}\mathbf{1}_K^T \mathbf{W} - \frac{1}{K}\mathbf{1}_K^T \mathbf{C}\mathbf{A}^T = \overline{\mathbf{w}}^T - \overline{\mathbf{c}}^T \mathbf{A}^T \tag{33}$$

where $\overline{\mathbf{c}} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{c}_i$ and $\overline{\mathbf{w}} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{w}_i$ are the centroids. Substituting this back into the objective function and solving for $\mathbf{A}$:

$$\frac{\partial \epsilon}{\partial \mathbf{A}} = -2\mathbf{C}^T (\mathbf{W} - \mathbf{C}\mathbf{A}^T - \mathbf{1}_K \mathbf{t}_c^T) = 0 \tag{34}$$

After centering the data by defining $\mathbf{C}' = \mathbf{C} - \mathbf{1}_K\bar{\mathbf{c}}^T$ and $\mathbf{W}' = \mathbf{W} - \mathbf{1}_K\overline{\mathbf{w}}^T$, the optimal affine transformation matrix is

$$\mathbf{A}^T = (\mathbf{C}'^T\mathbf{C}')^{-1}\mathbf{C}'^T\mathbf{W}' \tag{35}$$

The translation vector is then

$$\mathbf{t}_c = \overline{\mathbf{w}} - \mathbf{A}\bar{\mathbf{c}} \tag{36}$$

Finally, we construct the homogeneous transformation matrix that incorporates the general affine transformation

$$\mathbf{M}_c = \begin{bmatrix} \mathbf{A} & \mathbf{t}_c \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} \tag{37}$$

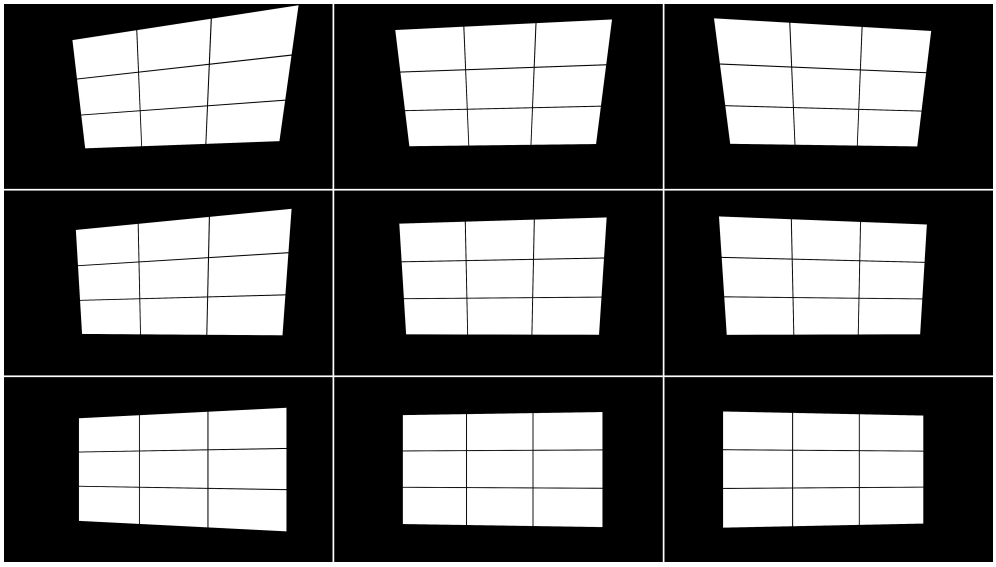
where $\mathbf{A} \in \mathbb{R}^{3\times3}$ is the unconstrained affine transformation matrix and $\mathbf{t}_c \in \mathbb{R}^3$ is the translation vector. This matrix represents a general affine transformation that maps any point from the camera coordinate system to the light field world coordinate system. Unlike rigid body transformations, this approach allows for arbitrary linear transformations including rotation, scale, skew, and other geometric distortions without any orthogonality or determinant constraints. By employing direct least squares regression, we minimize the calibration error and ensure that the transformation between the RGB-D sensor and the light field coordinate system accommodates all possible systematic distortions. Based on the solved projection matrix $\mathbf{M}_c$, we can obtain the position of the eye in the light field coordinate system $\mathbf{P}_e \in \mathbb{R}^3$ as below:

$$\begin{bmatrix} \mathbf{P}_e \\ 1 \end{bmatrix} = \mathbf{M}_c \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} \tag{38}$$

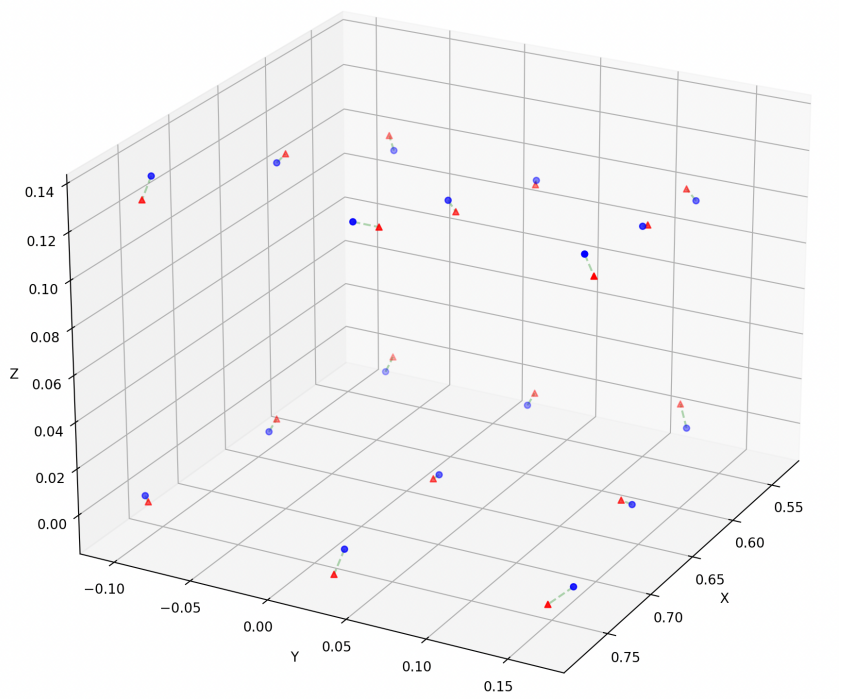
Here x_c, y_c, z_c represent the ocular coordinates in the RGB-D camera coordinate system.

4.2 Calibration experimental details and results

The calibration experiment was conducted to validate the theoretical framework for mapping eye-camera coordinates to light field world coordinates. To rigorously evaluate the proposed framework, synthetic data were generated using a physically accurate model to simulate retinal imaging under diverse geometric configurations. This approach ensured controlled conditions for assessing the robustness and accuracy of the coordinate transformation. The experimental setup was designed to systematically explore the parameter space, with a focus on key geometric variables that influence the mapping process. Specifically, radial distances $R \in \{40, 50, 60\}$ cm were selected to represent the distance from the physical world viewpoint to the world center, while polar angles $\phi \in \{70°, 80°, 90°\}$ and azimuthal angles $\theta \in \{80°, 90°, 100°\}$ were chosen to define the orientation of the viewpoint relative to the z-axis and y-axis, respectively. These parameter ranges were carefully selected to cover a broad



Supplementary Figure 4 Calibration charts for viewpoint match. Here are the front-depth right-eye calibration charts generated for $R = 40$ cm, with polar angles $\phi \in \{70°, 80°, 90°\}$ and azimuthal angles $\theta \in \{80°, 90°, 100°\}$. These calibration charts are used to validate the correspondence between physical world coordinates $\mathbf{c}_i$ and pixel world coordinates $\mathbf{w}_i$. The calibration image consists of a 10-pixel black line on a pure white background, positioned at the intersection of lines dividing the width and height into approximately three equal parts. This design facilitates the observer's judgment of the 3D effect.



Supplementary Figure 5 The visualization of calibration error. The plot illustrates the reprojection errors in both X, Y, and Z coordinates (in meters) across different calibration points. Here the errors consistently fall within acceptable limits for practical applications (<5cm), showcasing the simplicity but effectiveness of the proposed calibration method.

spectrum of potential viewpoints, ensuring comprehensive validation of the framework under varying observational conditions.

To facilitate the observer's judgment of the 3D effect, we designed a specialized calibration image. This image featured a 10-pixel black line drawn on a pure white background, positioned at the intersection of lines dividing the width and height into approximately three equal parts. As shown in Supplementary Figure 4, the first layer of calibration charts was generated for $R = 40$ cm, with ϕ values of $70°, 80°, 90°$ and θ values of $80°, 90°, 100°$. Using the model, we obtained calibration charts corresponding to different viewpoint coordinates. When these calibration charts were displayed on the screen, observers perceived a clear three-dimensional effect at the physical world coordinates $\mathbf{c}_i$. Since the pixel world coordinates $\mathbf{w}_i$ of the viewpoint were known, we employed the least squares method to solve for the transformation matrix $\mathbf{M}_c$ from the physical world to the pixel world. This approach provided an approximate solution for $\mathbf{M}_c$. Despite the simplicity of the proposed calibration method, the results demonstrated remarkable accuracy. As shown in Supplementary Figure 5, the reprojection errors were consistently low across all calibration points, and the angular deviations in the rotation component were negligible.

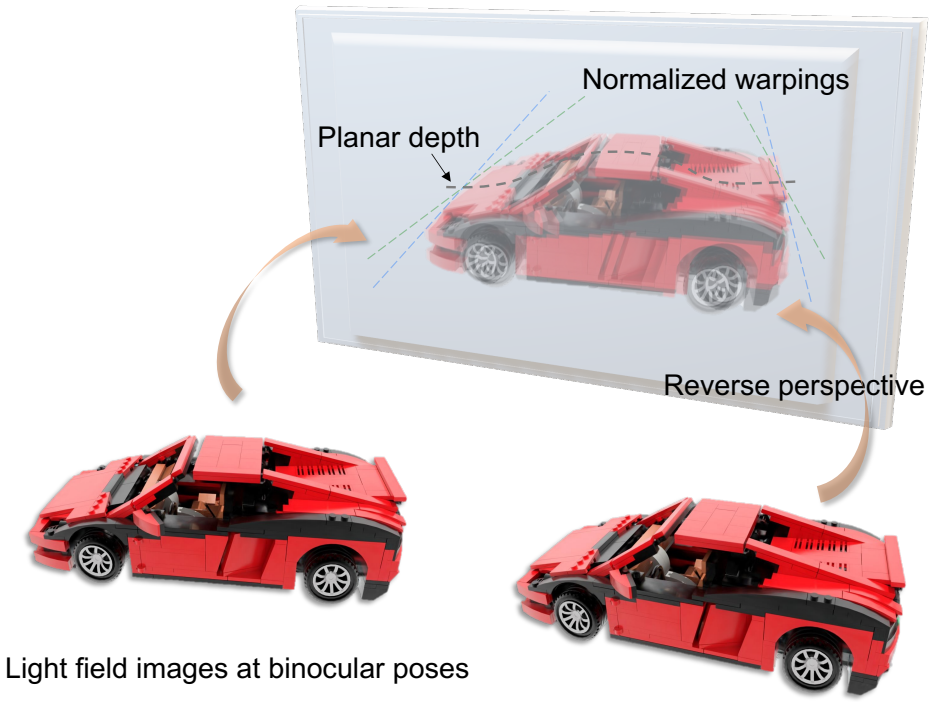
5 Ocular Geometric Encoding

We propose a Reverse Perspective Uniform (RPU) to encode the pose information of binocular views. To achieve a unified encoding, RPU employs reverse perspective transformation to standardize each image in binocular 6D poses onto the screen plane at each depth as normalized warpings (Supplementary Figure 6). The perspective transformation maps a three-dimensional scene onto a two-dimensional plane, simulating depth and spatial relationships within images. This transformation enables the image plane (viewing plane) to rotate about the perspective axis by a specific angle. This rotation modifies the original projection rays while preserving the geometric integrity of the shapes projected onto the image plane.

Mathematically, perspective transformation is typically represented by a homography matrix, a 3×3 matrix used to map points from the original image coordinates (j, k) to the transformed image coordinates (x, y). This transformation matrix can be decomposed into four components, each representing a different linear transformation, such as scaling, shearing, rotation, and translation, which collectively produce the perspective effect. These linear transformations can be considered special cases of perspective transformation, where the projection plane and the original plane are parallel, resulting in a simpler affine transformation. The general transformation equation for perspective transformation is expressed as

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} j \\ k \\ 1 \end{bmatrix} \tag{39}$$

Here, (j, k) are the coordinates in the original image, and (x', y') are the coordinates in the transformed image. The w' is a scale factor that accounts for the depth of the



Supplementary Figure 6 Illustration of the proposed Reverse Perspective Uniform (RPU) for encoding binocular pose information. The RPU applies reverse perspective transformations to project the binocular 6D poses onto the screen plane as normalized warpings. This transformation standardizes images from binocular camera coordinates onto the light field plane at a unified depth, preserving geometric integrity.

point relative to the camera. To obtain the final 2D coordinates (x, y), we normalize by dividing x' and y' by w':

$$\begin{cases} x = \dfrac{x'}{w'} \\ y = \dfrac{y'}{w'} \end{cases} \tag{40}$$

To determine the transformation matrix, it is necessary to have at least four corresponding point pairs between the original and the transformed images. Since the homography matrix has eight degrees of freedom (excluding the scale factor), these four point pairs provide eight linear equations, allowing the unique determination of the matrix elements. Consider the specific case of transforming a square into an arbitrary quadrilateral. Let the four corners of the square be mapped to four points on the quadrilateral. These points can be denoted as (j_1, k_1), (j_2, k_2), (j_3, k_3), and (j_4, k_4) in the original image, and (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4) in the transformed image. The transformation equations for these points are:

$$\begin{cases} x_1 = \dfrac{h_{11}j_1 + h_{12}k_1 + h_{13}}{h_{31}j_1 + h_{32}k_1 + h_{33}} \\ y_1 = \dfrac{h_{21}j_1 + h_{22}k_1 + h_{23}}{h_{31}j_1 + h_{32}k_1 + h_{33}} \\ \vdots \\ \vdots \\ x_4 = \dfrac{h_{11}j_4 + h_{12}k_4 + h_{13}}{h_{31}j_4 + h_{32}k_4 + h_{33}} \\ y_4 = \dfrac{h_{21}j_4 + h_{22}k_4 + h_{23}}{h_{31}j_4 + h_{32}k_4 + h_{33}} \end{cases} \tag{41}$$

In the light field display system discussed in the main text, the light field is oriented towards the eyes, thereby forming a retinal image. Here, the coordinates (j_i, k_i) correspond to the pixel coordinates on the light field plane, while the coordinates (x_i, y_i) correspond to the pixel coordinates on the eye camera. Specifically, (j_i, k_i) is denoted as $\mathbb{Q}_n := \{(0,0), (W,0), (W,H), (0,H)\}$ in the main text, while (x_i, y_i) is represented as $\mathbb{Q}'_n := \{(u'_i, v'_i)\}|_{i=1}^{4}$ in the main text. As aforementioned, here we exclude the scale

factor and set $h_{33} = 1$ for convenience. Substituting all these into Equation 41 yields

$$\begin{aligned}
u'_0 &= h_{13} \\
v'_0 &= h_{23} \\
u'_1 &= \frac{h_{11}W + h_{13}}{h_{31}W + 1} \\
v'_1 &= \frac{h_{11}W + h_{13}}{h_{31}W + 1} \\
u'_2 &= \frac{h_{11}W + h_{12}H + h_{13}}{h_{31}W + h_{32}H + 1} \\
v'_2 &= \frac{h_{21}W + h_{22}H + h_{23}}{h_{31}W + h_{32}H + 1} \\
u'_3 &= \frac{h_{12}H + h_{13}}{h_{32}H + 1} \\
v'_3 &= \frac{h_{22}H + h_{23}}{h_{32}H + 1}
\end{aligned} \tag{42}$$

Rearranging the above equation gives

$$\begin{aligned}
h_{13} &= u'_0 \\
h_{23} &= v'_0 \\
Wh_{11} + h_{13} - Wu'_1h_{31} - u'_1 &= 0 \\
Wh_{21} + h_{23} - Wv'_1h_{31} - v'_1 &= 0 \\
Wh_{11} + Hh_{12} + h_{13} - Wu'_2h_{31} - Hu'_2h_{32} - u'_2 &= 0 \\
Wh_{21} + Hh_{22} + h_{23} - Wv'_2h_{32} - Hv'_2h_{32} - v'_2 &= 0 \\
Hh_{12} + h_{13} - Hu'_3h_{32} - u'_3 &= 0 \\
Hh_{22} + h_{23} - Hv'_3h_{32} - v'_3 &= 0
\end{aligned} \tag{43}$$

To simplify the solution and make the relationships more compact, auxiliary variables can be introduced. With the auxiliary variables, the equations governing the transformation can be reduced, making it easier to understand and solve.

$$\begin{aligned}
\Delta u'_1 &= u'_1 - u'_2 \\
\Delta v'_1 &= v'_1 - v'_2 \\
\Delta u'_2 &= u'_3 - u'_2 \\
\Delta v'_2 &= v'_3 - v'_2 \\
\Delta u'_3 &= u'_0 - u'_1 + u'_2 - u'_3 \\
\Delta v'_3 &= v'_0 - v'_1 + v'_2 - v'_3
\end{aligned} \tag{44}$$

When these auxiliary variables are set to zero, the transformation plane becomes parallel to the original plane. This effectively reduces the transformation to an affine

transformation, which is a simpler type that does not account for perspective effects, where the size of objects changes based on their distance from the viewer. When they are non-zero, the result is a full perspective transformation, and the size, shape, and position of objects in the transformed image will vary based on their relative position to the viewer. Continuing to substitute these auxiliary variables and simplify the above equation can yield:

$$
\begin{aligned}
h_{31} &= \frac{\begin{vmatrix} \Delta u'_3 & \Delta u'_2 \\ \Delta v'_3 & \Delta v'_2 \end{vmatrix}}{W \begin{vmatrix} \Delta u'_1 & \Delta u'_2 \\ \Delta v'_1 & \Delta v'_2 \end{vmatrix}} \\
h_{32} &= \frac{\begin{vmatrix} \Delta u'_1 & \Delta u'_3 \\ \Delta v'_1 & \Delta v'_3 \end{vmatrix}}{H \begin{vmatrix} \Delta u'_1 & \Delta u'_2 \\ \Delta v'_1 & \Delta v'_2 \end{vmatrix}} \\
h_{21} &= \frac{v'_1 - v'_0}{W} + h_{31} v'_1 \\
h_{22} &= \frac{v'_3 - v'_0}{H} + h_{32} v'_3 \\
h_{11} &= \frac{u'_1 - u'_0}{W} + h_{31} u'_1 \\
h_{12} &= \frac{u'_3 - u'_0}{H} + h_{32} u'_3 \\
h_{13} &= u'_0 \\
h_{23} &= v'_0 \\
h_{33} &= 1
\end{aligned}
\tag{45}
$$

The above derives matrix elements of the forward perspective process. For the sake of binocular pose encoding, i.e., mapping the eye pose to the light field coordinate system, we can get the reverse perspective transformation matrix $\mathbf{M}_{RPU}$ from its inverse matrix given $\mathbb{Q}'_n$ of one eye:

$$
\mathbf{M}_{RPU} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}^{-1} \tag{46}
$$

6 Specific Settings of Lightfield Data Establishment

This section outlines the experimental configurations used for constructing our light field datasets. As summarized in Supplementary Tables 3 and 4, we document the key parameters necessary for reproducibility and cross-comparison, including the world-to-light-field scaling factor s, the compensating transformation matrix $\mathbf{M}_{\text{comp}}$ (with explicit rotation and translation), and the physical scene depth $d$thick defining the volumetric extent.

LightField Dataset	s	d_{thick}(cm)
uCO3D objects	0.21	6
MatrixCity	0.23	5
Shoe Rack	0.08	12
Hot Dog	0.06	10
Orchids	0.04	100
Red Car	0.16	6
Ficus	0.05	10
Chair	0.06	9.1
Desk & Bookshelf	0.03	6
Truck	0.08	6
Toy Dozer	0.08	11
Room Floor	0.02	46.7
Wukang Mansion	0.07	6
China Art Museum	0.03	25

Supplementary Table 3 Configuration parameters of light field datasets. The table presents the scaling factor s for physical-to-digital transformation and the corresponding depth d_{thick} in centimeters for each dataset.

LightField Dataset	**Rotation**	**Translation**
uCO3D objects	[0, 0, 0]	[0, 0, 0]
MatrixCity	[0, 0, 0]	[0, -1.5, 0]
Shoe Rack	[234, 0, 122]	[2.6, -0.24, 3.1]
Hot Dog	[0, -92, 14]	[2.9, 0.04, 0.16]
Orchids	[-90, 0, 90]	[19.16, -1.78, 0]
Red Car	[0, 0, 0]	[-1, 1.2, -1.6]
Ficus	[0, 0, 0]	[0, 0, 0]
Chair	[0, 0, 0]	[0, 0, 0]
Desk & Bookshelf	[-114, 0, 90]	[4.8, -2.5, 0]
Truck	[80, 182, -92]	[0.5, -0.24, 0.1]
Toy Dozer	[0, 0, 0]	[0, 0, 0]
Room Floor	[-20, -90, 86]	[0, 0.13, -3.07]
Wukang Mansion	[194, 23, -31]	[0.34, -0.02, 2.02]
China Art Museum	[234, 4, -61]	[0.54, 0.33, 0.72]

Supplementary Table 4 The transformation parameters for light field datasets. The table lists the rotation (in degrees) and translation (in meters) components of the compensating matrix for each dataset.

References

[1] Tay, S., Blanche, P.-A., Voorakaranam, R., Tunç, A., Lin, W., Rokutanda, S., Gu, T., Flores, D., Wang, P., Li, G., *et al.*: An updatable holographic three-dimensional display. Nature **451**(7179), 694–698 (2008)

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

[3] Shi, L., Li, B., Matusik, W.: End-to-end learning of 3d phase-only holograms for holographic display. Light: Science & Applications **11**(1), 247 (2022)

[4] Park, J., Lee, K., Park, Y.: Ultrathin wide-angle large-area digital 3d holographic display using a non-periodic photon sieve. Nature communications **10**(1), 1304 (2019)

[5] Dodgson, N.A.: Variation and extrema of human interpupillary distance. In: Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 36–46 (2004). SPIE

[6] Wan, W., Qiao, W., Huang, W., Zhu, M., Ye, Y., Chen, X., Chen, L.: Multiview holographic 3d dynamic display by combining a nano-grating patterned phase plate and lcd. Optics express **25**(2), 1114–1122 (2017)

[7] Nam, D., Lee, J.-H., Cho, Y.H., Jeong, Y.J., Hwang, H., Park, D.S.: Flat panel light-field 3-d display: concept, design, rendering, and calibration. Proceedings of the IEEE **105**(5), 876–891 (2017)

[8] Van Berkel, C., Clarke, J.A.: Characterization and optimization of 3d-lcd module design. In: Stereoscopic Displays and Virtual Reality Systems IV, vol. 3012, pp. 179–186 (1997). SPIE

[9] Takaki, Y., Nago, N.: Multi-projection of lenticular displays to construct a 256-view super multi-view display. Optics express **18**(9), 8824–8835 (2010)

[10] Konrad, J., Halle, M.: 3-d displays and signal processing. IEEE Signal Processing Magazine **24**(6), 97–111 (2007)

[11] Lanman, D., Hirsch, M., Kim, Y., Raskar, R.: Content-adaptive parallax barriers: optimizing dual-layer 3d displays using low-rank light field factorization. In: ACM SIGGRAPH Asia 2010 Papers, pp. 1–10 (2010)

[12] Kawakita, M., Iwasawa, S., Sakai, M., Haino, Y., Sato, M., Inoue, N.: 3d image quality of 200-inch glasses-free 3d display system. In: Stereoscopic Displays and Applications XXIII, vol. 8288, pp. 63–70 (2012). SPIE

[13] Yu, H., Yan, X., Yan, Z., Chen, Z., Liu, J., Jiang, X.: Analysis and enhancement

for ultra-large depth-of-field light field display based on mems laser scanning projection arrays. Optics Express **32**(20), 34898–34919 (2024)

[14] Li, X., Ding, J., Zhang, H., Chen, M., Liang, W., Wang, S., Fan, H., Li, K., Zhou, J.: Adaptive glasses-free 3d display with extended continuous viewing volume by dynamically configured directional backlight. OSA continuum **3**(6), 1555–1567 (2020)

[15] Li, H., Wang, S., Zhao, Y., Chen, S., Li, T.: Integral imaging display method based on holographic diffuser and discrete lens array. In: Optoelectronic Imaging and Multimedia Technology VII, vol. 11550, pp. 53–65 (2020). SPIE

[16] Zhang, H.-L., Ma, X.-L., Lin, X.-Y., Xing, Y., Wang, Q.-H.: System to eliminate the graininess of an integral imaging 3d display by using a transmissive mirror device. Optics Letters **47**(18), 4628–4631 (2022)

[17] Kooi, F.L., Toet, A.: Visual comfort of binocular and 3d displays. Displays **25**(2-3), 99–108 (2004)

[18] Akiduki, H., Nishiike, S., Watanabe, H., Matsuoka, K., Kubo, T., Takeda, N.: Visual-vestibular conflict induced by virtual reality in humans. Neuroscience letters **340**(3), 197–200 (2003)

[19] Wetzstein, G., Lanman, D., Hirsch, M., Heidrich, W., Raskar, R.: Compressive light field displays. IEEE computer graphics and applications **32**(5), 6–11 (2012)

[20] Gotoda, H.: A multilayer liquid crystal display for autostereoscopic 3d viewing. In: Stereoscopic Displays and Applications XXI, vol. 7524, pp. 227–234 (2010). SPIE

[21] Wetzstein, G., Lanman, D., Heidrich, W., Raskar, R.: Layered 3d: tomographic image synthesis for attenuation-based light field and high dynamic range displays. In: ACM SIGGRAPH 2011 Papers, pp. 1–12 (2011)

[22] Zhu, L., Lv, G., Xv, L., Wang, Z., Feng, Q.: Performance improvement for compressive light field display based on the depth distribution feature. Optics Express **29**(14), 22403–22416 (2021)

[23] Maruyama, K., Takahashi, K., Fujii, T.: Comparison of layer operations and optimization methods for light field display. IEEE Access **8**, 38767–38775 (2020)

[24] Sun, Y., Li, Z., Li, L., Wang, S., Gao, W.: Optimization of compressive light field display in dual-guided learning. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2075–2079 (2022). IEEE

[25] Sun, Y., Li, Z., Wang, S., Gao, W.: Depth-assisted calibration on learning-based

factorization for a compressive light field display. Optics Express **31**(4), 5399–5413 (2023)

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

[27] Dodgson, N.A.: 3d without the glasses. Nature **495**, 316–317 (2013)

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

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

[30] Fan, Z.-B., Qiu, H.-Y., Zhang, H.-L., Pang, X.-N., Zhou, L.-D., Liu, L., Ren, H., Wang, Q.-H., Dong, J.-W.: A broadband achromatic metalens array for integral imaging in the visible. Light: Science & Applications **8**(1), 67 (2019)

[31] Greivenkamp, J.E.: Field guide to geometrical optics. (2004). SPIE Bellingham

[32] Gopakumar, M., Lee, G.-Y., Choi, S., Chao, B., Peng, Y., Kim, J., Wetzstein, G.: Full-colour 3d holographic augmented-reality displays with metasurface waveguides. Nature, 1–7 (2024)

[33] Yu, H., Lee, K., Park, J., Park, Y.: Ultrahigh-definition dynamic 3d holographic display by active control of volume speckle fields. Nature Photonics **11**(3), 186–192 (2017)

[34] Dorrah, A.H., Bordoloi, P., Angelis, V.S., Sarro, J.O., Ambrosio, L.A., Zamboni-Rached, M., Capasso, F.: Light sheets for continuous-depth holography and three-dimensional volumetric displays. Nature Photonics **17**(5), 427–434 (2023)

[35] Li, J., Smithwick, Q., Chu, D.: Holobricks: modular coarse integral holographic displays. Light: Science & Applications **11**(1), 57 (2022)

[36] Wang, D., Li, Z.-S., Zheng, Y., Zhao, Y.-R., Liu, C., Xu, J.-B., Zheng, Y.-W., Huang, Q., Chang, C.-L., Zhang, D.-W., *et al.*: Liquid lens based holographic camera for real 3d scene hologram acquisition using end-to-end physical model-driven network. Light: Science & Applications **13**(1), 62 (2024)

[37] Qin, Y., Chen, W.-Y., O'Toole, M., Sankaranarayanan, A.C.: Split-lohmann multifocal displays. ACM Transactions on Graphics (TOG) **42**(4), 1–18 (2023)

[38] Lv, G.-J., Wang, Q.-H., Zhao, W.-X., Wang, J.: 3d display based on parallax barrier with multiview zones. Applied Optics **53**(7), 1339–1342 (2014)

[39] Stern, A., Javidi, B.: Three-dimensional image sensing, visualization, and processing using integral imaging. Proceedings of the IEEE **94**(3), 591–607 (2006)

[40] Martinez-Cuenca, R., Saavedra, G., Martinez-Corral, M., Javidi, B.: Progress in 3-d multiperspective display by integral imaging. Proceedings of the IEEE **97**(6), 1067–1077 (2009)

[41] Xia, X., Zhang, X., Zhang, L., Surman, P., Zheng, Y.: Time-multiplexed multi-view three-dimensional display with projector array and steering screen. Optics Express **26**(12), 15528–15538 (2018)

[42] Pamplona, V.F., Oliveira, M.M., Aliaga, D.G., Raskar, R.: Tailored displays to compensate for visual aberrations. ACM transactions on graphics (TOG) **31**(4), 1–12 (2012)

[43] Huang, F.-C., Wetzstein, G., Barsky, B.A., Raskar, R.: Eyeglasses-free display: towards correcting visual aberrations with computational light field displays. ACM transactions on graphics (TOG) **33**(4), 1–12 (2014)

[44] Fattal, D., Peng, Z., Tran, T., Vo, S., Fiorentino, M., Brug, J., Beausoleil, R.G.: A multi-directional backlight for a wide-angle, glasses-free three-dimensional display. Nature **495**(7441), 348–351 (2013)

[45] Hwang, Y.S., Bruder, F.-K., Fäcke, T., Kim, S.-C., Walze, G., Hagen, R., Kim, E.-S.: Time-sequential autostereoscopic 3-d display with a novel directional backlight system based on volume-holographic optical elements. Optics Express **22**(8), 9820–9838 (2014)

[46] Watanabe, H., Kawakita, M., Okaichi, N., Sasaki, H., Kano, M., Mishina, T.: High-resolution spatial image display with multiple uhd projectors. In: Three-Dimensional Imaging, Visualization, and Display 2018, vol. 10666, pp. 180–188 (2018). SPIE

[47] Watanabe, H., Okaichi, N., Omura, T., Kano, M., Sasaki, H., Kawakita, M.: Aktina vision: Full-parallax three-dimensional display with 100 million light rays. Scientific reports **9**(1), 17688 (2019)

[48] Lanman, D., Wetzstein, G., Hirsch, M., Heidrich, W., Raskar, R.: Polarization fields: dynamic light field display using multi-layer lcds. In: Proceedings of the 2011 SIGGRAPH Asia Conference, pp. 1–10 (2011)

[49] Huang, F.-C., Chen, K., Wetzstein, G.: The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. ACM Trans. Graph. **34**(4) (2015)

[50] Maimone, A., Wetzstein, G., Hirsch, M., Lanman, D., Raskar, R., Fuchs, H.: Focus 3d: Compressive accommodation display. ACM Trans. Graph. **32**(5), 153–1 (2013)

[51] Ebner, C., Mohr, P., Langlotz, T., Peng, Y., Schmalstieg, D., Wetzstein, G.,

Kalkofen, D.: Off-axis layered displays: Hybrid direct-view/near-eye mixed reality with focus cues. IEEE transactions on visualization and computer graphics **29**(5), 2816–2825 (2023)

[52] Davis, J.A., McNamara, D.E., Cottrell, D.M., Sonehara, T.: Two-dimensional polarization encoding with a phase-only liquid-crystal spatial light modulator. Applied Optics **39**(10), 1549–1554 (2000)

[53] Moreno, I., Martínez, J.L., Davis, J.A.: Two-dimensional polarization rotator using a twisted-nematic liquid-crystal display. Applied optics **46**(6), 881–887 (2007)

[54] Wetzstein, G., Lanman, D., Hirsch, M., Raskar, R.: Tensor displays: compressive light field synthesis using multilayer displays with directional backlighting. ACM transactions on graphics (TOG) **31**(4) (2012)

[55] Hirsch, M., Lanman, D., Wetzstein, G., Raskar, R.: Construction and calibration of optically efficient lcd-based multi-layer light field displays. In: Journal of Physics: Conference Series, vol. 415, p. 012071 (2013). IOP Publishing

[56] Wang, S., Liao, W., Surman, P., Tu, Z., Zheng, Y., Yuan, J.: Salience guided depth calibration for perceptually optimized compressive light field 3d display. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2031–2040 (2018)

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

[58] Chen, D., Sang, X., Yu, X., Zeng, X., Xie, S., Guo, N.: Performance improvement of compressive light field display with the viewing-position-dependent weight distribution. Optics express **24**(26), 29781–29793 (2016)

[59] Kim, Y.-D., Choi, S.: Weighted nonnegative matrix factorization. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1541–1544 (2009). IEEE

[60] Zhang, J., Fan, Z., Sun, D., Liao, H.: Unified mathematical model for multilayer-multiframe compressive light field displays using lcds. IEEE transactions on visualization and computer graphics **25**(3), 1603–1614 (2018)

[61] Cichocki, A., Zdunek, R., Amari, S.-i.: Nonnegative matrix and tensor factorization [lecture notes]. IEEE signal processing magazine **25**(1), 142–145 (2007)

[62] Clark Jones, R.: A new calculus for the treatment of optical systems. Journal of the Optical Society of America (1942)

[63] Collett, E.: Field guide to polarization (2005). Spie Bellingham

[64] Goodman, J.W.: Introduction to Fourier Optics. McGraw-Hill physical and quantum electronics series, (2005)

[65] Yeh, P., Gu, C.: Optics of liquid crystal displays **67** (2009)

[66] Zhu, L., Chen, Q., Chen, T., Lv, G., Feng, Q., Wang, Z.: High-brightness hybrid compressive light field display with improved image quality. Optics Letters **48**(23), 6172–6175 (2023)