Holographic Parallax Improves 3D Perceptual Realism

DONGYEON KIM* and SEUNG-WOO NAM*, Seoul National University, Republic of Korea
SUYEON CHOI*, Stanford University, USA
JONG-MO SEO, Seoul National University, Republic of Korea
GORDON WETZSTEIN, Stanford University, USA
YOONCHAN JEONG, Seoul National University, Republic of Korea

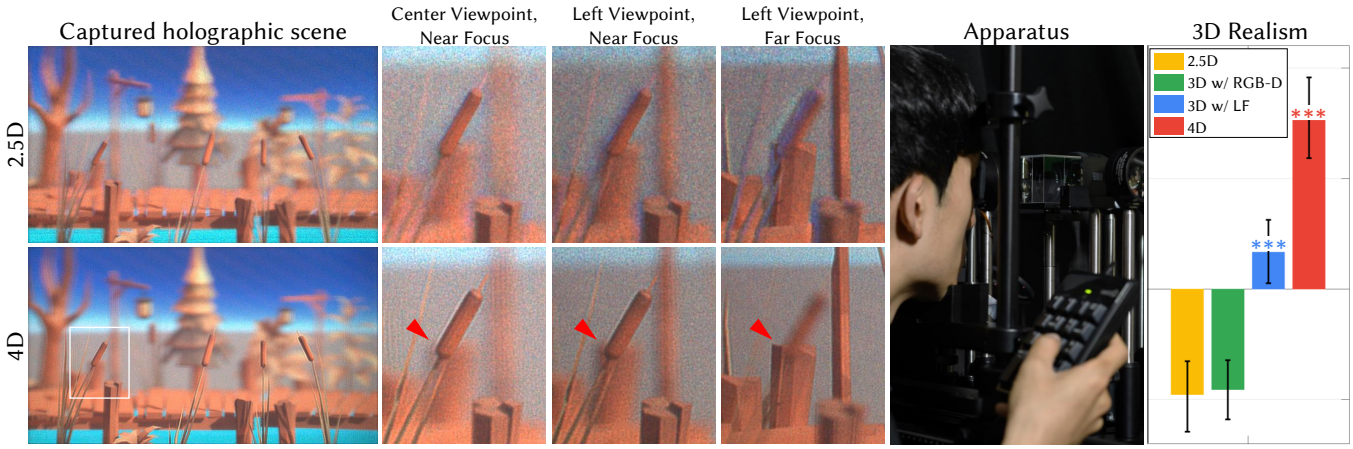

Fig. 1. **Experimental results from our perceptual holographic testbed** (*left*). With recent advancements in computer-generated hologram (CGH) algorithms, the image quality of holographic displays has surpassed the threshold required for conducting robust user studies, enabling us to investigate the perceptual impact brought by modern holographic displays. While a significant amount of effort has been made for specific target formats in ideal viewpoints, they lack to deliver correct parallax cues under natural viewing conditions (*top row*). In contrast, light field holograms successfully convey these cues (*bottom row*) highlighted with red arrows in the enlarged images provided with different capture settings (position, focus). We build a full-color, high-quality holographic testbed, where we conduct a user study examining 3D perceptual realism across various 3D CGH algorithms (*right*). Our results reveal that CGH algorithms designed for specific types of targets significantly lag in perceptual realism, whereas the light field hologram notably outperforms other formats. Asterisks (red: 4D over the other formats, blue: 3D w/ LF over 2.5D and 3D w/ RGB-D) indicate the statistical significance of the difference (***: p<0.001), and the errorbars denote 95% confidence interval. (purchased Unity asset: Low Poly Series: Landscape)

Holographic near-eye displays are a promising technology to solve long-standing challenges in virtual and augmented reality display systems. Over the last few years, many different computer-generated holography (CGH) algorithms have been proposed that are supervised by different types of target content, such as 2.5D RGB-depth maps, 3D focal stacks, and 4D light fields. It is unclear, however, what the perceptual implications are of the choice of algorithm and target content type. In this work, we build a perceptual testbed of a full-color, high-quality holographic near-eye display. Under natural viewing conditions, we examine the effects of various CGH supervision formats and conduct user studies to assess their perceptual impacts on 3D realism. Our results indicate that CGH algorithms designed for specific viewpoints exhibit noticeable deficiencies in achieving 3D realism. In contrast, holograms incorporating parallax cues consistently outperform other formats across different viewing conditions, including the center of the eyebox. This finding is particularly interesting and suggests that the inclusion of parallax cues in CGH rendering plays a crucial role in enhancing the overall quality of the holographic experience. This work represents an initial stride towards delivering a perceptually realistic 3D experience with holographic near-eye displays.

CCS Concepts: • **Hardware** → **Emerging technologies**.

Additional Key Words and Phrases: virtual reality, augmented reality, computational displays, holography, perception

ACM Reference Format:
Dongyeon Kim, Seung-Woo Nam, Suyeon Choi, Jong-Mo Seo, Gordon Wetzstein, and Yoonchan Jeong. 2024. Holographic Parallax Improves 3D Perceptual Realism. *ACM Trans. Graph.* 43, 4, Article 68 (July 2024), 13 pages. https://doi.org/10.1145/3658168

*Authors contributed equally to this research.

Authors' Contact Information: Dongyeon Kim, dongyeon93@snu.ac.kr; Seung-Woo Nam, 711asd@snu.ac.kr, Seoul National University, Republic of Korea; Suyeon Choi, suyeon@stanford.edu, Stanford University, USA; Jong-Mo Seo, callme@snu.ac.kr, Seoul National University, Republic of Korea; Gordon Wetzstein, gordon.wetzstein@stanford.edu, Stanford University, USA; Yoonchan Jeong, yoonchan@snu.ac.kr, Seoul National University, Republic of Korea.

ACM 1557-7368/2024/7-ART68
https://doi.org/10.1145/3658168

1 INTRODUCTION

Holographic displays offer great potential as the next-generation platform for augmented and virtual reality displays [Jang et al.,

2024, Maimone et al., 2017] due to their versatile functionalities providing high-resolution volumetric images with aberration and vision correction [Kim et al., 2021] capabilities arising from complex amplitude modulation of light. Nevertheless, achieving superior holographic visualization through the utilization of spatial light modulators (SLMs) that operate solely on either phase or amplitude has been a long-standing challenge in the field. Recently, several significant breakthroughs in holographic image quality incorporating machine learning-based approaches have shown a promising path toward a renaissance of computational holography [Chakravarthula et al., 2020, Nam et al., 2023, Peng et al., 2020, Shi et al., 2021, Yang et al., 2022]. However, most evaluations of these holographic displays have been conducted using camera-based experiments, which only consider specific viewing conditions.

Unlike stationary cameras, the human eye is constantly in motion, involuntarily experiencing pupil contractions and relaxation [Bahill et al., 1975]. In contrast to incoherent displays, which have developed gaze-contingent approaches to address eye movement [Guan et al., 2022, Mercier et al., 2017], holographic displays possess unique capabilities in controlling the plenoptic function of light [Choi et al., 2022, Park, 2020]. However, the limited size of the eyebox [Jang et al., 2018] and computational load of holographic displays often leads to approximating the 3D scene based on the center view and overlooking the impact on other views [Shi et al., 2021].

These two facts pose a number of fundamental questions to the field of holographic displays: Does an approximated holographic 3D scene, optimized for a camera or an ideal viewpoint, truly provide a satisfying 3D experience for users? How robust are these approximations, and are the perceived differences discernible even within the current optical settings of holographic near-eye displays with limited étendue? Secondly, if we aim to determine the optimal format for high-quality 3D holographic scenes by addressing the aforementioned question, what criteria must be met in order to surpass perceptual thresholds?

In this study, we investigate the perceptual realism of 3D scenes presented through holographic near-eye displays, while considering natural viewing conditions. Our approach includes simulating the perceptual quality of the 3D holographic scenes with varying computer-generated holography (CGH) target content types such as 2.5D RGB-depth maps, 3D focal stacks and 4D light fields, and pupil conditions and accounting for the impact of eye movements on sampled signals. To find the best CGH supervision format for realistic 3D holographic scenes, we conduct user studies. Our results show that incorporating parallax cues significantly enhances the 3D user experience, even with limited head movement. This study represents a first step in the field of 3D visual experience with holographic near-eye displays and provides guidelines for creating perceptually realistic 3D holographic scenes. (See Fig. 1)

The contributions of this study are as follows.

- We simulate the impacts of eye movement, pupil size fluctuation, and directional sensitivity of the retina on the perceived 3D holographic scenes, which implies discrepancies between camera-based experiments and evaluations involving humans.
- We design and conduct user studies under various viewing conditions to determine the optimal formats that holographic displays need to reproduce in order to achieve 3D perceptual realism. To this end, we build a perceptual testbed of a holographic near-eye display with high-quality, full-color 3D holographic scenes.
- The user studies reveal the findings indicating that the 3D CGH supporting parallax cues significantly improves 3D perceptual realism in various viewing conditions, even with limited head movements.

2 RELATED WORK

2.1 Computer-generated holography

Computer-generated holography (CGH) encompasses algorithms that generate holograms for spatial light modulators (SLMs), manipulating the complex-valued incident wave field to achieve desired light field distributions for viewers. These algorithms have been developed to accommodate various 3D data formats, including image layers [Shi et al., 2022, Zhang et al., 2017], RGB-D [Chen et al., 2021, Choi et al., 2021, Shi et al., 2021], focal stacks [Choi et al., 2022, Kavaklı et al., 2023, Yang et al., 2022], or polygons [Matsushima and Nakahara, 2009, Wang et al., 2023], with essential occlusion handling [Symeonidou et al., 2015].

It is noteworthy that intensity-based data formats do not inherently impose constraints on the phase distribution of holograms, introducing uncertainty to their plenoptic function. One can assume a random phase to simulate diffused light [Lohmann and Paris, 1967], or a smooth phase which might result in better contrast [Maimone et al., 2017, Shi et al., 2021] but at the cost of a reduced eyebox size [Choi et al., 2022].

Recent studies have emphasized the tradeoff between image quality and the eyebox. Yoo et al. [2021] investigated controlling randomness to strike a balance, and stochastic pupil sampling can ensure consistent 2D image appearance across the eyebox [Chakravarthula et al., 2022]. However, these approaches have limitations in accurately expressing spatial-angular information across the étendue [Kuo et al., 2020, Park et al., 2019]. Light field holograms, also known as holographic stereograms [Choi et al., 2022, Kang et al., 2016, Padmanaban et al., 2019a, Park, 2020, Shi et al., 2017, Zhang et al., 2015], present a promising solution to address this challenge.

2.2 Visual perception

Virtual reality (VR) systems aim to offer immersive experiences by understanding the human visual system and perceptual studies build guidelines for the relevant communities. Recent advancements in visual difference predictors (VDP) [Mantiuk et al., 2021] are used to estimate perceived image quality, considering different display configurations and observance models.

Depth perception mechanisms are vital for evaluating the perceptual realism of 3D scenes. Various cues, including binocular disparity, accommodation, convergence, and motion parallax, contribute to depth perception [Cutting and Vishton, 1995]. Aligning these cues is crucial to reduce visual fatigue [Hoffman et al., 2008] during prolonged VR device use. Images with binocular disparity are most sensitive to human perception, providing the primary cue for depth

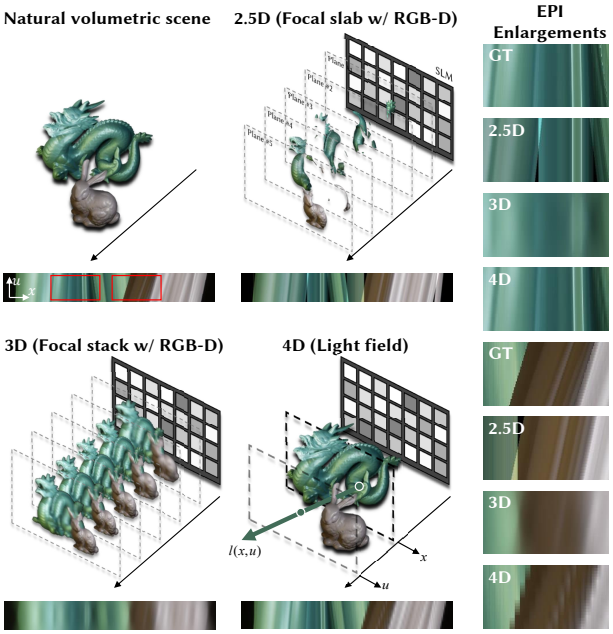

Fig. 2. **Various CGH supervision targets** (2.5D, 3D, 4D) **for holographic displays to realize the natural volumetric scene** (ground truth, GT). The reconstructed epipolar plane images (EPIs) of the individual data formats are provided to demonstrate the angle-dependent spatial information and the red-boxed regions are enlarged to demonstrate the differences. The EPIs are reconstructed with 25 horizontal views for the GT case, 5 planar images for 2.5D and 3D cases, and 5 view images for 4D case. Dragon, Bunny: credit to Stanford Computer Graphics Laboratory.

perception within arm's reach. Motion parallax, perceived through retinal motion, is the strongest depth cue supported for objects approximately one meter or more away. Retinal motion-driven depth perception is aided by the smooth pursuit eye movement [Naji and Freeman, 2004, Nawrot, 2003]. Rendering VR scenes, taking into account ocular parallax [Konrad et al., 2020]—the change in viewpoint due to eye rotation—has improved perceptual realism. However, the evaluation was done with stereo 3D head-mounted displays.

Assessing 3D realism can be subjective, incorporating various cues for image and depth perception, relying on individual visual behavior. However, the primary aim of VR displays persists in achieving *3D perceptual realism* and successfully passing the visual Turing test [Wetzstein and Lanman, 2016], particularly when assessing volumetric scenes under natural viewing conditions. Recent studies [March et al., 2022, Zhong et al., 2021] have conducted visual Turing tests using dual-plane stereo displays, representing progress in advancing next-generation display technologies.

2.3 Perceptual 3D holographic testbed

Conducting meaningful user studies with holographic displays has historically been challenging due to low image quality, characterized by speckles and imperfect representation of the complex-valued field, resulting in low contrast compared to other displays. However, recent advancements in CGH and SLMs have significantly improved image quality through techniques like time-multiplexing and calibration [Chakravarthula et al., 2020, Choi et al., 2022, Curtis et al., 2021, Lee et al., 2020, 2022, Peng et al., 2020]. These advancements enable more accurate and robust user studies with holographic displays. Kim et al. [2022] conducted a user study with holographic near-eye displays, enhancing accommodation response using CGH supervision with a regularizer on the contrast ratio of two-dimensional (2D) images. However, the study focused solely on 2D content and did not consider parallax cues. For a comparison of 3D perceptual testbeds, please refer to Table S1.

Table 1. **Assessment of data formats in terms of supported visual cues**. In 2.5D and 3D formats, the blur behavior is constrained by the phase profile or the pupil size used in rendering the focal stack. Conversely, the defocus behavior in light fields would be accurate, though blur size could be limited by the étendue supported by the display system. The light field format and supervision uniquely enable the display to render correct view-dependent effects, such as occlusion, parallax, and specular highlights.

	multiple points in a single ray	retinal blur	view dependency
2.5D	no	approx.	approx.
3D w/ RGB-D	yes	approx.	approx.
3D w/ LF	yes	correct	approx.
4D	yes	correct	correct

2.4 Visual effects from 3D assets

The visual experience of a display device, especially for 3D content, depends on the presented content. The epipolar plane image (EPI) in Fig. 2 represents a horizontal cross-section image of ray-space ($l(x,u)$), defined with space (x) and direction (u). Different data formats have limitations: focal slabs (2.5D target) lack spatial information in angles other than the normal angle ($u = 0$) [Chang et al., 2020]. Focal stacks (3D) may overfit to the central view (3D w/ RGB-D) but can be generated with multiple views (3D w/ LF). Light field (4D) offers angle-dependent spatial information but with sparse sampling. The intrinsic nature of the light field enables the reproduction of angle-dependent visual effects such as occlusion and shading. For a comparison of visual effects supported by individual assets, please refer to Table 1.

3 UNIFORM 3D HOLOGRAPHIC EXPERIENCE ACROSS THE EYEBOX

Different 3D data formats possess inherent visual effects, and as the dimensions increase, so does the computational complexity. This raises the question of whether 3D near-eye displays necessitate the reconstruction of a 4D light field rather than relying on approximated 3D information since the perceived view-dependent effects highly vary on viewing conditions. Major aspects that affect the experience of view-dependent visuals include the eyebox size of the 3D near-eye display and the pupil status of the human eye.

In this section, our goal is to determine the most suitable 3D data format for visualizing perceptually realistic 3D scenery and assess it using a holographic near-eye display capable of rendering various 3D assets. To ensure comprehensive results, we conduct simulations

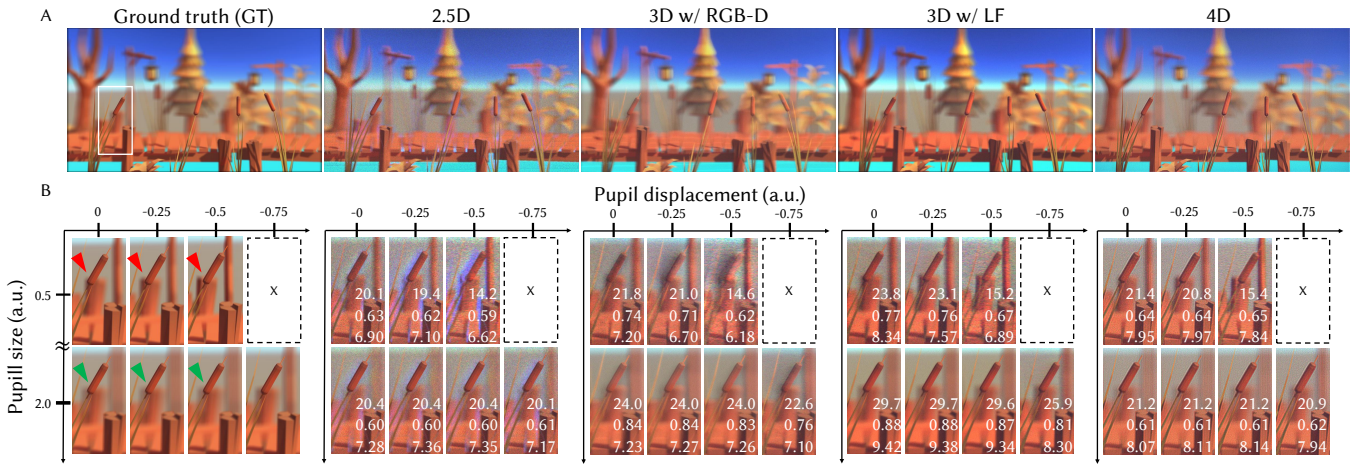

Fig. 3. **Holographic reconstruction with different CGH supervision targets:** (A) Near-depth focused holographic images (2.5D (2nd col, focal slab), 3D w/ RGB-D (3rd col, focal stack with RGB-D), 3D w/ LF (4th col, focal stack with 25×25 LF), 4D (5th col, 9×9 LF)) of the landscape_day scene are reconstructed in the full eyebox condition, respectively with the ground truth (GT) focal stack (1st col). (B) Enlargements of the corresponding holographic scenes reconstructed based on 7 different pupil settings (pupil displacement and size) are presented except the one with the fully vignetted (box with dashed line) condition. Each enlargement is consecutively provided with the quality metrics of PSNR, SSIM (maximum of 1), FovVideoVDP (JOD unit having a maximum of 10) [Mantiuk et al., 2021] evaluated with the GT focal stack. Here, the pupil displacement ($x_{p,norm}$) presents the eye pupil's displacement (x_p) in the horizontal axis and the pupil size ($D_{p,norm}$) denotes the diameter of the human eye pupil (D_p), and those values are normalized with the width of the eyebox (w_{eyebox}, 2.2 mm). The enlargements with red arrows indicate scenes reconstructed under an overfilled pupil and those with green arrows denote images visualized under an underfilled pupil. (purchased Unity asset: Low Poly Series: Landscape)

and experiments across different scenarios involving variations in eyebox and pupil status.

3.1 Simulation

In practice, the eye rotates to gaze at the objects located across the field of view leading to pupil displacement. Moreover, the pupil varies in size depending on the intensity of light entering it, and there is significant variation in pupil size among individuals. The holographic images were reconstructed to examine the impact of different pupil states (displacement and size) as shown in Fig. 3. As addressed by previous works [Chakravarthula et al., 2022, Kim et al., 2022] concerning the eyebox of holographic near-eye displays, the human eye pupil, while not considering its focal state, located in the eyebox domain optically acts as a binary low-pass filter and it samples the display signal.

The holographic images are reconstructed in different pupil states and evaluated with different quality metrics (peak signal-to-noise (PSNR), structural similarity metric (SSIM), FovVideoVDP [Mantiuk et al., 2021]) as provided in Fig. 3. Here, we included the FovVideoVDP as it outperformed other quality metrics in terms of evaluation of light field dataset through stereo 3D displays [Kiran Adhikarla et al., 2017]. The ground truth focal stack is generated with a dense light field (25×25 views) and processed with the identical pupil state. In detail, FovVideoVDP (v1.2.0) is estimated in a non-foveated mode with the condition of 86.2 [pix/deg], Lpeak=100, Lblack=0.1 [cd/m^2] and the results are scaled in a unit of Just-Objectionable-Difference (JOD). Note that the difference of 1 JOD refers to the visual difference that 75 percent of subjects choose the option compared to the counterpart and serves as the perceptual difference threshold.

When comparing the supervision of 4D CGH with center-view based CGH supervision (2.5D and 3D w/ RGB-D) using conventional image metrics like PSNR or SSIM, it is observed that 4D-supervised CGH results in relatively similar or sometimes poorer quality across the eyebox. However, the reconstructed results show better FovViodeVDP exceeding around 1 JOD, as described in Fig. 3(B). The assessment using FovVideoVDP ensures the improved perceptual quality of 4D across the eyebox.

3.1.1 Eyebox-pupil scenarios. Although determining the field of view and the size of the eyebox is one of the design considerations for holographic near-eye displays with limited étendue, the pupil size of the human eye fluctuates based on luminance. This results in various scenarios regarding the ratio of the exit-pupil and ocular-pupil areas [Ratnam et al., 2019]. Based on this relation in size, we can categorize into two major eyebox-pupil scenarios: an overfilled pupil when the size of the human eye pupil is smaller than the eyebox and an underfilled pupil when the size of the eye pupil is larger than the eyebox.

In the overfilled pupil scenarios (pupil states indicated by red arrows in Fig. 3(B)), the quality of the 3D w/ LF case deteriorates as the pupil is decentered while 4D shows smaller falloffs. If there is a difference of about 1 JOD in the comparison of the cases (3D w/ LF vs. 4D) in the specific pupil state ($(x_{p,norm}, D_{p,norm})$ =(-0.5, 0.5)), it may affect the quality of the overall viewing experience with eye movements. We additionally provide simulation and experimental results captured with the display system in Supplementary Material.

In contrast to the overfilled pupil scenarios, the 3D w/ LF case outperforms the 4D case in underfilled pupil scenarios (pupil states with green arrows in Fig. 3(B)) in terms of the simulated metrics. It is worth noting that evaluation with perceptual metrics may underestimate the impact of human factors and the influence of ocular parallax on perceptual realism since the metrics are built upon 2D image-based assessment. Therefore, in the subsequent subsection, we conduct a user study under the worst-case scenario, where the eyebox is significantly smaller compared to the eye pupil, evaluating various 3D data formats within this context.

3.2 User evaluation

3.2.1 Hardware and software. We built a benchtop prototype of a holographic near-eye display with a single SLM having a resolution of 1920 (H) × 1200 (V), and a full-color laser diode as a perceptual testbed for user validation as Fig. 4(A). In addition, we additionally placed an eye tracker to adjust the eye position of participants and simultaneously record the subjects' pupil position and size. The prototype provides an image with a resolution of 1600 × 900, and the corresponding field of view of 18.6° × 10.5°. The maximum resolution achieved by the system is 43 cycle per degree (cpd). The eyebox of the near-eye display defined with the blue illumination and a 40-mm focal length eyepiece lens is 2.2 mm × 1.1 mm as we place a side-band filter to modulate a complex-valued field with amplitude-encoded CGH. We carefully designed the perceptual testbed, ensuring that its maximum resolution surpasses the human visual acuity of 30 cpd [Guenter et al., 2012] while maintaining an eyebox size smaller than the average human eye pupil [De Groot and Gebhard, 1952].

The SLM utilized in the user experiment supports the full-color speckle-reduced image with temporal multiplexing of 24 binary CGHs. The CGH acquisition with various target formats is implemented with Pytorch based on the previous works of the differentiable time-multiplexed CGH optimization frameworks [Choi et al., 2022, Lee et al., 2022]. Detailed information on software and hardware can be found in Appendix and Supplementary Material.

3.2.2 Stimuli and Conditions. For the user validation, three volumetric scenes - landscape_day, landscape_night, and village - are used as stimuli as presented in Fig. 4(B). The depth range extends from 0 diopter (*D*) to 9.57 *D*, spanning from optical infinity to 11 cm from the eye. This range sufficiently covers the average accommodation range of young adults [Duane, 1912], and ocular parallax induced by the objects with the given depth range exceeds the minimal angular resolution of the fovea region [Konrad et al., 2020]. The display scheme is explained in Fig. S1. The luminance of each scene is estimated as 2 cd/m^2, and the room was kept dark during the experiment.

In the case of 2.5D and 3D-supervised CGHs, nine planes equally spaced in a unit of diopter are sampled. For 3D-supervised CGHs, we prepared two scenarios; *3D w/ RGB-D* and *3D w/ LF*. For *3D w/ RGB-D*, the focal stacks are generated with a single RGB-D map blending occlusion boundaries [Lee et al., 2022]. For *3D w/ LF*, we utilized LF with 25×25 orthographic views to generate focal stacks to naturally handle occlusion. Lastly, LF with 9×9 orthographic

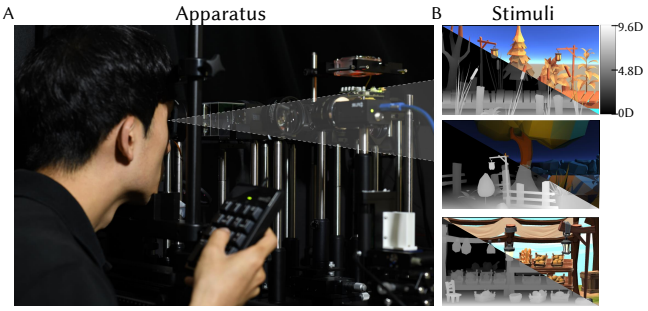

Fig. 4. **3D holographic perceptual testbed and stimuli.** (A) We conduct the user study using the apparatus shown on the left. Holographic scenes are generated using various CGH methods, using targets rendered with scenes purchased from the Unity Asset Store (Low Poly Series: Landscape, Fantastic-Village Pack)

views is utilized for 4D CGH supervision [Choi et al., 2022]. The captured scenes of each stimulus are provided in Fig. 5 and Fig. S16.

The experiment is done with four different viewing conditions; *Center* refers to the case when the subjects view 3D contents while placing the pupil at the 'sweet spot' of the eyebox and this represents the underfilled pupil. *Decentered* and *Vignetted* refer to the condition when the eye is horizontally decentered about 1.25 mm and 2.5 mm from the center, respectively. *w/ head movement* refers to the viewing condition when the subjects perform the task without head movement restriction. Note that none of the viewing conditions limited eye movements. We referred to the viewing conditions as *Center*, *Decentered*, *Vignetted* to differentiate the conditions based on the initial placement of the eye.

Before the experiment, six complete pairs with four different CGH supervision cases (*2.5D*, *3D w/ RGB-D*, *3D w/ LF*, *4D*) were prepared, the order was randomly shuffled to eliminate the potential decision bias and each pair was repeatedly provided three times. The complete pairwise comparison was held with three different scenes (landscape_day, landscape_night, and village) in four different viewing conditions (*Center*, *Decentered*, *Vignetted*, *w/ head movement*). The whole number of trials was 216 (6 pairs × 3 repetitions × 3 scenes × 4 viewing conditions).

3.2.3 Subjects. We recruited 28 naïve participants under the age of 40 (ranging from 23 to 36 with an average of 27.6, 12 female) to account for the potential decrease in accommodation range with age. All participants had normal or corrected-to-normal vision and normal color vision. They were rewarded for their participation, and the studies adhered to the Declaration of Helsinki. All subjects provided voluntary written and informed consent, and the experiment was conducted after receiving approval from the Institutional Review Board of the host institution.

3.2.4 Procedure. Before each session, precise head alignment was performed. Subjects were instructed to restrict their head movement in viewing conditions other than *w/ head movement*. The head position of the subjects was controlled by adjusting the components of the chin-and-head rest. During this adjustment procedure, subjects were asked to maintain their gaze over the center object of the

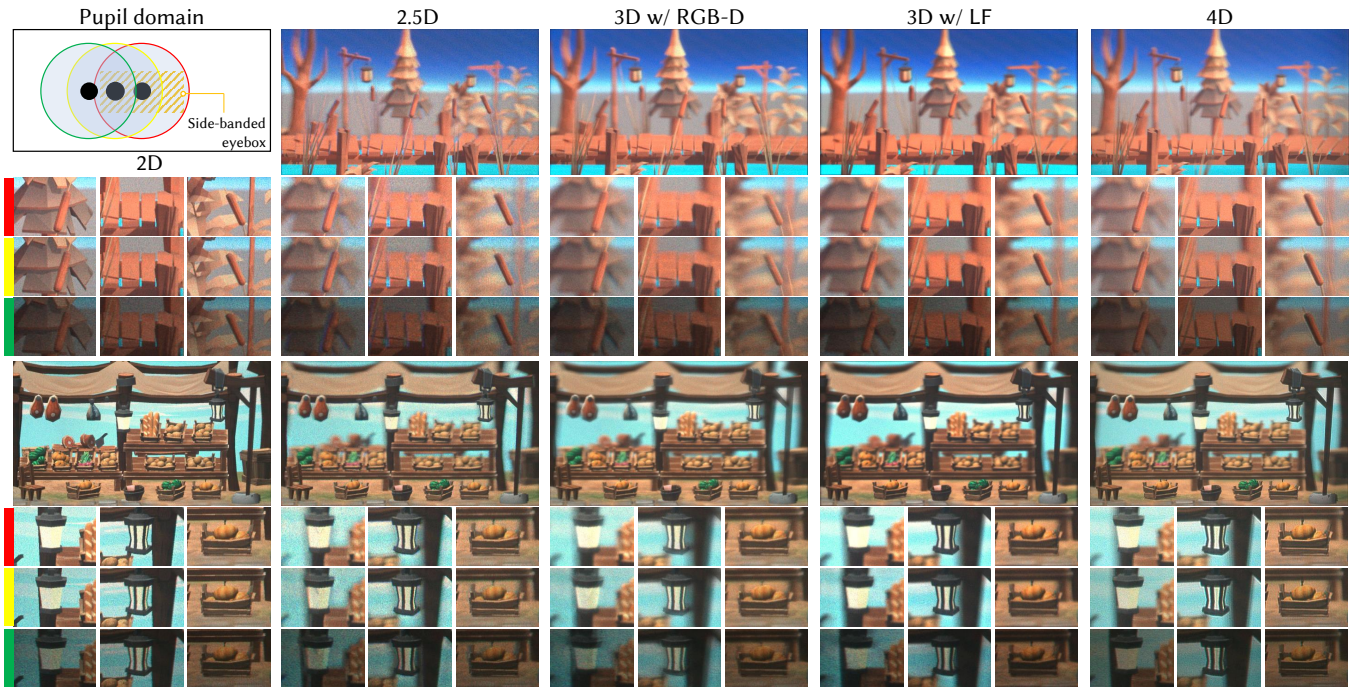

Fig. 5. **Experimental results with different pupil positions.** Holographic scenes supervised with 2D (1st col), 2.5D (2nd), 3D w/ RGB-D (3rd), 3D w/ LF (4th), 4D (5th) targets are captured with different pupil positions (red: $(x_{p,norm}, D_{p,norm}) = (0, 1.1)$, yellow: $(-0.68, 1.1)$ and green: $(-1.36, 1.1)$). The scenes are photographed with different focal states (landscape_day: 7th, village: 7th) out of 9 distinct focal states equally sampled in diopter. Enlargements are provided with the image focused on the magnified object. The colors of each row for the enlargements indicate the pupil positions (red: center, yellow: decentered, green: vignetted). We intentionally provide the results without modifying the brightness to show the energy across the eyebox. Note that it is hard to discriminate 3D w/ LF case and 4D case with the captured results. (purchased Unity asset: Low Poly Series: Landscape, Fantastic-Village Pack)

sample scene, utilizing eye-tracked data monitored in real-time. For the viewing condition *w/ head movement*, the chin-and-head rest was removed, and subjects were free to move their heads within a range where scenes remained observable.

In every viewing condition, a two-interval forced choice (2-IFC) [Bogacz et al., 2006] task was conducted, asking subjects to choose the 'more realistic 3D' option after presenting a pair of stimuli in sequence. Subjects were instructed to gaze at different objects and aim for a sharp focus on the gazed object to assess 3D quality, eliminating subjects who maintained focus at a single plane. Each pair of stimuli was displayed for 8 seconds, with a second of a gray noisy image provided in between. Responses were recorded using a keypad, and the next pair was presented after a valid input. After each session, subjects were encouraged to take a break for at least a minute, and the entire experiment took around an hour.

3.2.5 Results. The CGHs supervised with different targets were evaluated in four distinct viewing conditions and compared in terms of perceived 3D realism, as depicted in Fig. 6(A). The accumulated vote counts from a total of 26 subjects were normalized and scaled using the unit of JOD. The responses of two subjects were excluded after outlier analysis introduced by the work of Perez-Ortiz and Mantiuk [2017].

Upon conducting the two-tailed z-test with the scaled JOD scores for each CGH supervision target in each viewing condition, the results indicate that 4D-supervised CGHs exhibit significant improvements in perceived 3D quality across all viewing conditions compared to other forms of CGH supervision. Especially compared with *2.5D* and *3D w/ RGB-D*, the difference exceeds 1 JOD in some conditions. Additionally, *3D w/ LF* is significantly preferred over *2.5D* and *3D w/ RGB-D*, and this preference is even more pronounced in the viewing condition involving head movement. Notably, no significant differences were observed between *2.5D* and *3D w/ RGB-D* in any of the viewing conditions.

In summary, the results demonstrate significant differences in 3D perceptual realism in every viewing condition when the parallax cues were taken into account in CGH supervision. The 4D-supervised CGH outperformed all other cases by considerable margins, and even the *3D w/ LF* CGH outperformed the other cases with strong evidence of significance.

Throughout the experiment, the position and size of the subjects' pupils were monitored and recorded. Figure 6(B) presents the measured data of one representative subject in a single session, with different colors indicating different viewing conditions. The measured data demonstrate that the experiments were carried out under various viewing positions. Furthermore, it is noteworthy that even when head movement was restricted, the eye exhibited substantial movement.

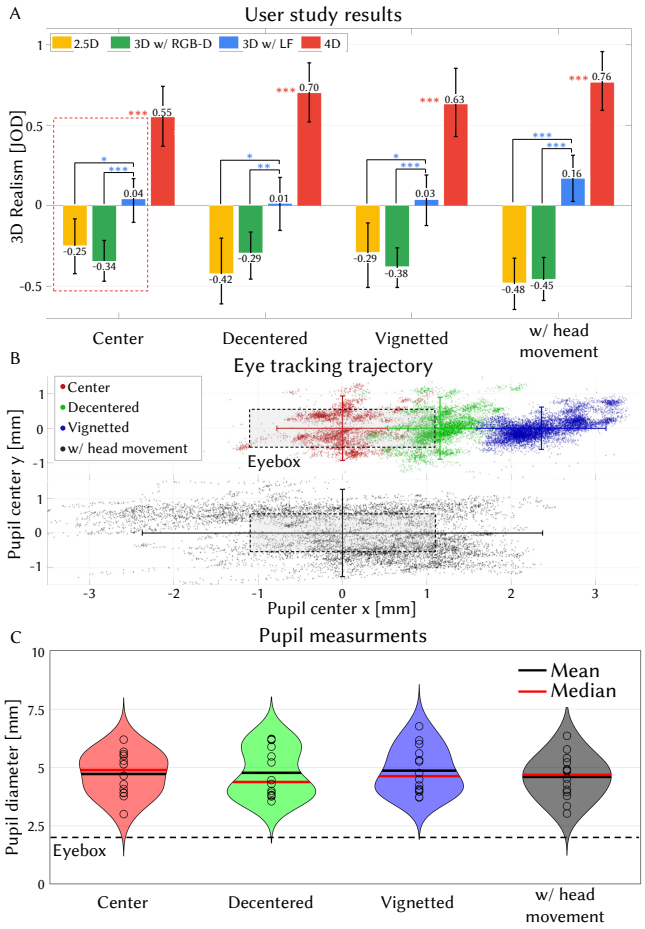

Fig. 6. **User experiment results:** (A) 3D realism is assessed using CGHs supervised with four target formats (2.5D in yellow, 3D w/ RGB-D in green, 3D w/ LF in blue, 4D in red) across four viewing conditions (Center, Decentered, Vignetted, and with head movement). The mean JOD is set as zero for each viewing condition. Error bars represent 95% confidence intervals estimated by bootstrapping 500 samples. Asterisks (blue: 3D w/ LF vs. paired cases, red: 4D vs. other cases) indicate the statistical significance of differences (*: p<0.05, **: p<0.01, ***: p<0.001). (B) The tracked trajectory of the pupil center for one representative subject. Error bars represent the 95% confidence interval of the pupil displacement. (C) Measured pupil diameters of representative subjects depending on the viewing conditions. The black circle corresponds to the pupil diameter of individual subjects and the dashed line denotes the width of eyebox in our experimental setup.

We provide the pupil diameter measured data with the equipped eye-tracker depending on four different viewing conditions as shown in Fig. 6(C). We excluded the data corrupted by eye blinking by subjecting the measured data achieving the confidence level of 0.85. The average pupil diameters were measured from 4.5 mm to 5 mm. This measured value exceeds pupil diameter of 4.4 mm ($D_{p,norm} = 2.0$) ensuring that most of the pupil positions recorded in the viewing condition of *Center* as the underfilled pupil condition. Due to the imperfect pupil diameter measurement with the eye-tracker, the measured data of 14 subjects are presented. The mean pupil diameter of 5 mm corresponds to the case when the luminance is low around 1 cd/m^2 [Napieralski and Rynkiewicz, 2019]. If the luminance level is as high as the level supported by the conventional VR displays (hundreds of nits) [Mehrfard et al., 2019], the pupil diameter would be smaller and the impact of parallax cues will magnify in the overfilled pupil conditions as observed in Fig. 3.

It is intriguing that even within the experiment conducted at the *Center* viewing condition that the pupil is large enough to cover the eyebox, the *4D* approach outperforms the *3D w/ LF* approach in terms of the perceived quality of 3D visuals. Note the retinal image of human eye is based on the focal stack. Notably, the focal stacks in the *3D w/ LF* case are generated with 25×25 views, while *4D* employs 9×9 orthographic views for CGH supervision. Although the VDP estimated in the specific viewing condition is 0.51 JOD which does not exceed 1 JOD value, the discrepancy between the simulated VDP and the actual VDP from user studies could potentially open up a vast research field in 3D quality metrics. This reversal in preference will be further discussed in the discussion section.

4 HOW MANY LIGHT FIELD VIEWS ARE REQUIRED?

The prior evaluation emphasized the substantial improvement in 3D realism with 4D CGH supervision, especially supporting holographic parallax over other CGH supervision methods in diverse viewing conditions. Then, exploring the ideal number of views for 4D CGH supervision is crucial for efficient rendering, considering the holographic testbed's adaptable specifications—a feature not commonly available in other 3D displays.

4.1 Experimental results

Fig. 7 presents the experimental results captured with the benchtop prototype of a holographic near-eye display, illustrating the impact of the number of views used in 4D CGH supervision on 3D visualization. Additional reconstructed holographic scenes, varying based on the number of views, can be found in Fig. S12-S13.

4.2 User study

4.2.1 Procedure. We obtained CGHs supervised with different view counts: 3×3, 5×5, 7×7, and 9×9. The optical setup, utilizing a sideband filter, resulted in effective view counts of 3×1, 5×2, 7×3, and 9×4, maintaining view gaps. Three scenes were used as stimuli in the initial experiment, and each pairwise comparison was repeated five times, totaling 90 trials lasting approximately thirty minutes. Eyebox centering was ensured before commencing the test, without recording eye-tracking data. All subjects that performed the first user experiment participated in the test and the overall procedure is identical to the first experiment.

Results. We conducted an evaluation comparing CGHs supervised using 4D targets with variations in the number of views. The responses of 24 subjects were analyzed, excluding four participants' responses after outlier analysis with JOD scores estimated from accumulated vote counts over scenes. After removing the outliers, we estimated the confidence interval using the bootstrapping method. The statistical test was conducted using a two-tailed z-test on the JOD scores obtained for each viewing condition. The results in

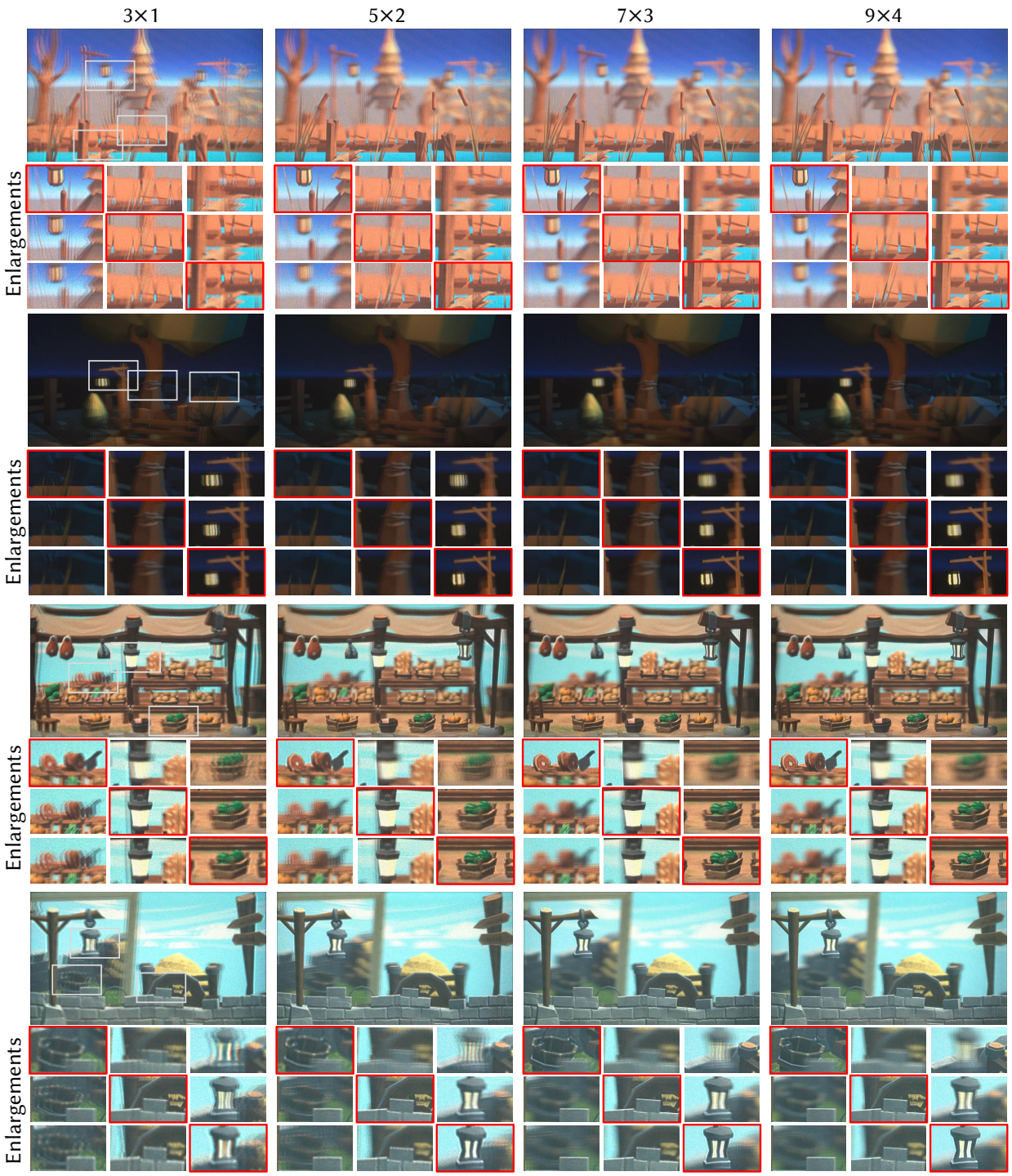

Fig. 7. **Experimental results with the different number of views used for 4D CGH supervision.** The 3D holographic scenes of the sampled focal states are provided. Insets show the images captured with three different focuses randomly chosen for each scene and the images in gray boxes are enlarged. Among the enlarged images, the red boxes indicate that the object is focused. (purchased Unity asset: Low Poly Series: Landscape, Fantastic-Village Pack)

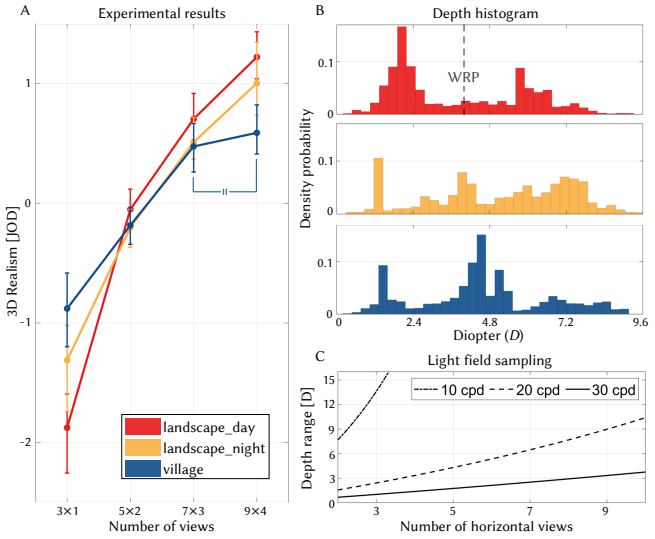

Fig. 8. **The effect of the number of views employed in 4D CGH supervision on the perceived 3D realism:** (A) The JOD-scaled results of the pairwise comparison are provided depending on the scene. The equal symbol indicates that a statistically significant difference is not observed between the paired conditions of the scene. The errorbars indicate the 95% confidence interval acquired with bootstrapping. (B) Depth histogram profiles of stimuli. The dashed line indicates the WRP's dioptric depth of the system. (C) The relationship between the number of horizontal views used in the 4D CGH supervision and the depth range expressible by the system is plotted with the three distinct spatial bandwidths (dotdash line: 10 cpd, dashed line: 20 cpd, solid line: 30 cpd) of the targets.

Fig. 8(A) were dependent on the number of views. Specifically, JOD values in each number of view conditions (3×1, 5×2, 7×3, 9×4) were scaled for three different scenes. Differences in 3D realism between neighboring view number conditions were evaluated by a two-tailed z-test with scaled JOD scores, and significant differences were observed in every paired case except for one (7×3 vs. 9×4 in the village scene, p=0.45).

Scene-specific findings regarding the results can be better understood by referring to the depth distribution in Fig. 8(B). For the village scene, objects are relatively concentrated near the depth of the wavefront recording plane (WRP) compared to other scenes. The overall depth range can be understood by considering the light field sampling theorem [Ng et al., 2005, Park and Askari, 2019]. If the depth range extends from 0 diopters to D_{max}, the overall depth range supported by the near-eye display, is depicted in a unit of diopter in Fig. 8(C). It depends on the scene's spatial bandwidth (Bx), as

$$D_{max} = \frac{2N_u}{f(2N_u - f\lambda B_x^2)}, \tag{1}$$

where, f represents the focal length of the eyepiece lens, λ is the wavelength of the light source, and N_u denotes the angular resolution. With the signal subjected to low-pass filtering, the depth range gets broader. However, if the scene extends to the spatial bandwidth of 30 cpd, more views are required to secure a certain depth range. Interestingly, the analysis of depth representation remains consistent with different optical specifications of the eyepiece lens as provided in Fig. S14.

5 DISCUSSION

We conducted user studies using a testbed of modern holographic near-eye displays to determine the ideal 3D formats for providing perceptual reality. Our findings revealed that reconstructing parallax across the eyebox realized with 4D light field as CGH supervision target enhances 3D perceptual realism across various eyebox scenarios in VR near-eye displays.

5.0.1 Directional sensitivity of human eye. The human eye exhibits greater sensitivity to light entering near the center of the pupil than to light near the edge, primarily due to the directional sensitivity of cone photoreceptors. This phenomenon, commonly referred to as the Stiles-Crawford effect [Westheimer, 2008], can be described by modulating the pupil apodization profile (A) as follows:

$$A(x_p, y_p) = A_o(x_p - x_c, y_p - y_c)10^{-p(\lambda)((x_p-x_c)^2+(y_p-y_c)^2)}. \tag{2}$$

Here, (x_p, y_p), (x_c, y_c) respectively represents the coordinates of the pupil domain and those of the pupil center in a meter scale. A_o stands for the original apodization function which is a circular and binary filter, and $p(\lambda)$ is a wavelength-dependent parameter representing the magnitude of the Stiles-Crawford effect. For simulation, we have chosen a constant value of $2.5 \cdot 10^4$ [Westheimer, 2008] for this parameter across all color channels, disregarding wavelength differences.

This characteristic of the human eye, unlike the camera, results in a nontrivial result in the eyebox scenario of an underfilled pupil shown as Fig. 9. Additional reconstructed results with different CGH algorithms assuming the apodized pupil can be found in Fig. S11. It is worth noting that observing parallax images is also valid even in the extremely large pupil condition as the apodization profile exponentially decreases with the displacement. Although there are individual differences in the optical characteristics of the human visual system, precise optical modeling of downstream optics would help in understanding the perceived image.

5.0.2 User study stimuli and apparatus. The validation could have been performed with low-level psychophysics methodologies [Watson and Pelli, 1983] to identify the concrete discrimination and detection thresholds for the various viewing parameters such as depth range and number of views. However, these methods require densely sampled evaluation sets with various light field sets and CGHs, which require excessive time and memory. Despite these challenges, our stimuli, comprising complex scenes with depth distribution, effectively elicited natural eye movements provided in Fig. S20. In addition, conducting perceptual studies using binocular holographic near-eye displays can yield more definitive results due to the combined influence of binocular retinal disparity and retinal motion. Finally, conducting VDP simulations with display model parameters (luminance and contrast) matching the actual experimental conditions can potentially improve predictions. However, minor calibration adjustments are unlikely to fully explain the reversal of perceptual realism reported in the experiment, as shown in Fig. S15. Low-level psychophysics experiments conducted on a precisely measured display testbed, which provides 'full' depth

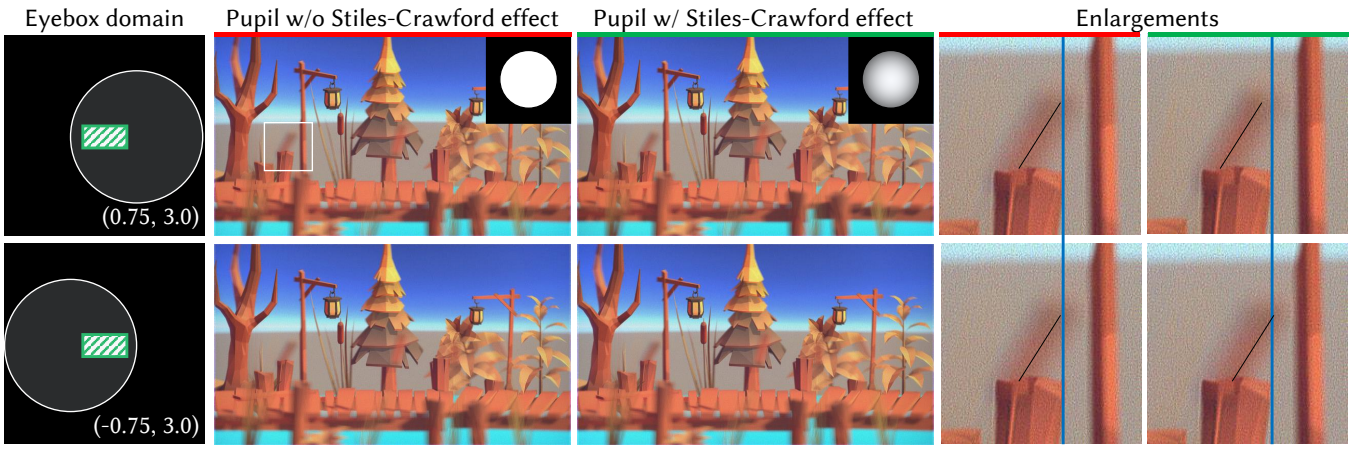

Fig. 9. **Reconstructed images with different pupil apodization profiles of the human eye** (2nd col: diffraction-limited pupil, 3rd col: apodized pupil considering Stiles-Crawford effect). The images are reconstructed with the 4D supervised CGH of the landscape_day scene when the center of the eye pupil largely deviates from the eyebox and the eye pupil is sufficiently large not to partially sample the eyebox as demonstrated in the (1st col) illustration of the eyebox domain. In the pupil case (1st row), the eye's pupil is decentered to its rightmost extent (a shift of 1.65 mm, $x_{p,norm} = 0.75$), while in the second scenario (2nd row), the eye is at its leftmost extent (a shift of -1.65 mm, $x_{p,norm} = -0.75$). This shift can be converted to the eye rotation of 9.37 degrees, which is almost equivalent to half of the horizontal FoV. In both cases, it is assumed that the eye's pupil is sufficiently large with a diameter of 6.6 mm ($D_{p,norm} = 3.0$) to cover the entire eyebox (2.2 mm × 1.1 mm). The pupil apodization profile is provided at the top right corner of each column. The identical part of the individual image is cropped and enlarged for better visualization. The blue line is drawn to represent the identical index of the horizontal plane, and the black line is drawn to better visualize the center of the defocused cattail of the scene. (purchased Unity asset: Low Poly Series: Landscape)

cues, can accelerate exploration into uncharted realms of human 3D perception.

5.0.3 Perceptual quality metric of 3D contents. There have been discrepancies between the realism predicted by the advanced perceptual quality metric [Mantiuk et al., 2021] and the user study results with 3D content. This can be attributed to the fact that the metrics are built upon 2D displays and conventional displays do not typically incorporate 3D visualization. Previous work on gaze-contingent ocular parallax VR rendering [Konrad et al., 2020] reported an ocular parallax detection threshold of ±0.36 *D* in eccentricity of 15°. With the given parameter, ocular parallax detection can be roughly analyzed as discussed in Sec. S5.1.5. This potential integration of perception thresholds in various domains presents an intriguing opportunity for researchers in optics, graphics, and vision science to explore perceptual metrics specifically tailored for evaluating 3D visual stimuli produced by modern displays.

5.0.4 Degrees of freedom, the number of constraints, and étendue. Improved perceptual realism achieved through light field optimization stems from the rich spatio-angular information provided by the target, which translates into an increased number of constraints. In our user study, we compared algorithms with the same limited degrees of freedom using our SLM, while there is a tight trade-off between the number of degrees of freedom, the number of constraints, and étendue [Monin et al., 2022a,b]. Here, we define the number of degrees of freedom as (number of optimizable pixels) × (number of frames), and the number of constraints as (number of pixels in the target) × (number of views) × (number of planes).

We perform additional simulations on these factors to verify the trend (see Fig. S24); as expected, increased degrees of freedom or a smaller number of constraints lead to low loss values. However, the commonly used mean square error metric should not directly represent the perceptual performance as mentioned in the previous subsection. Moreover, étendue expansion is another crucial direction as it would directly increase the possibility of eye displacement, likely magnifying our trend. From this perspective, our results hold significance as the study is conducted in a limited étendue setting (underfilled pupil).

6 CONCLUSION

Our work provides crucial insights on the effectiveness and realism of CGH algorithms that will help guide the community toward passing the visual Turing test of displays using future holographic light field near-eye displays.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00787, Development of vision assistant HMD and contents for legally blind and low visions). Suyeon Choi is supported by a Meta Research Ph.D. Fellowship and a Kwanjeong Scholarship.

REFERENCES

A. T. Bahill, M. R. Clark, and L. Stark. The main sequence, a tool for studying human eye movements. *Mathematical biosciences*, 24(3-4):191–204, 1975.

R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700, 2006.

O. Bryngdahl and A. Lohmann. Single-sideband holography. *JOSA*, 58(5):620–624, 1968.

P. Chakravarthula, E. Tseng, T. Srivastava, H. Fuchs, and F. Heide. Learned hardware-in-the-loop phase retrieval for holographic near-eye displays. *ACM Transactions on Graphics (TOG)*, 39(6):1–18, 2020.

P. Chakravarthula, S.-H. Baek, F. Schiffers, E. Tseng, G. Kuo, A. Maimone, N. Matsuda, O. Cossairt, D. Lanman, and F. Heide. Pupil-aware holography. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022.

J.-H. R. Chang, A. Levin, B. V. Kumar, and A. C. Sankaranarayanan. Towards occlusion-aware multifocal displays. *ACM Transactions on Graphics (TOG)*, 39(4):68–1, 2020.

C. Chen, B. Lee, N.-N. Li, M. Chae, D. Wang, Q.-H. Wang, and B. Lee. Multi-depth hologram generation using stochastic gradient descent algorithm with complex loss function. *Opt. Express*, 29(10):15089–15103, 2021.

S. Choi, M. Gopakumar, Y. Peng, J. Kim, and G. Wetzstein. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Trans. Graph. (SIGGRAPH Asia)*, 2021.

S. Choi, M. Gopakumar, Y. Peng, J. Kim, M. O'Toole, and G. Wetzstein. Time-multiplexed neural holography: a flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022.

V. R. Curtis, N. W. Caira, J. Xu, A. G. Sata, and N. C. Pégard. Dcgh: dynamic computer generated holography for speckle-free, high fidelity 3d displays. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 1–9. IEEE, 2021.

J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In *Perception of space and motion*, pages 69–117. Elsevier, 1995.

S. De Groot and J. Gebhard. Pupil size as determined by adapting luminance. *JOSA*, 42 (7):492–495, 1952.

A. Duane. Normal values of the accommodation at all ages. *Journal of the American Medical Association*, 59(12):1010–1013, 1912.

J. W. Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

P. Guan, O. Mercier, M. Shvartsman, and D. Lanman. Perceptual requirements for eye-tracked distortion correction in vr. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022.

B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder. Foveated 3d graphics. *ACM transactions on Graphics (tOG)*, 31(6):1–10, 2012.

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

C. Jang, K. Bang, G. Li, and B. Lee. Holographic near-eye display with expanded eye-box. *ACM Trans. Graph.*, 37(6), dec 2018.

C. Jang, K. Bang, M. Chae, B. Lee, and D. Lanman. Waveguide holography for 3d augmented reality glasses. *Nature Communications*, 15(1):66, 2024.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

H. Kang, E. Stoykova, and H. Yoshikawa. Fast phase-added stereogram algorithm for generation of photorealistic 3d content. *Applied optics*, 55(3):A135–A143, 2016.

K. Kavaklı, Y. Itoh, H. Urey, and K. Akşit. Realistic defocus blur for multiplane computer-generated holography. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 418–426. IEEE, 2023.

D. Kim, S.-W. Nam, K. Bang, B. Lee, S. Lee, Y. Jeong, J.-M. Seo, and B. Lee. Vision-correcting holographic display: evaluation of aberration correcting hologram. *Biomedical Optics Express*, 12(8):5179–5195, 2021.

D. Kim, S.-W. Nam, B. Lee, J.-M. Seo, and B. Lee. Accommodative holography: improving accommodation response for perceptually realistic holographic displays. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022.

V. Kiran Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk. Towards a quality metric for dense light fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 58–67, 2017.

R. Konrad, A. Angelopoulos, and G. Wetzstein. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Transactions on Graphics (TOG)*, 39(2):1–12, 2020.

G. Kuo, L. Waller, R. Ng, and A. Maimone. High resolution étendue expansion for holographic displays. *ACM Transactions on Graphics (TOG)*, 39(4):66–1, 2020.

B. Lee, D. Yoo, J. Jeong, S. Lee, D. Lee, and B. Lee. Wide-angle speckleless dmd holographic display using structured illumination with temporal multiplexing. *Optics Letters*, 45(8):2148–2151, 2020.

B. Lee, D. Kim, S. Lee, C. Chen, and B. Lee. High-contrast, speckle-free, true 3d holography via binary cgh optimization. *Scientific reports*, 12(1):2811, 2022.

A. W. Lohmann and D. Paris. Binary fraunhofer holograms, generated by computer. *Applied optics*, 6(10):1739–1748, 1967.

A. Maimone, A. Georgiou, and J. S. Kollin. Holographic near-eye displays for virtual and augmented reality. *ACM Trans. Graph. (SIGGRAPH)*, 36(4):85, 2017.

R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021.

J. March, A. Krishnan, S. Watt, M. Wernikowski, H. Gao, A. Ö. Yöntem, and R. Mantiuk. Impact of correct and simulated focus cues on perceived realism. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

K. Matsushima and S. Nakahara. Extremely high-definition full-parallax computer-generated hologram created by the polygon-based method. *Applied optics*, 48(34): H54–H63, 2009.

A. Mehrfard, J. Fotouhi, G. Taylor, T. Forster, N. Navab, and B. Fuerst. A comparative analysis of virtual reality head-mounted display systems. *arXiv preprint arXiv:1912.02913*, 2019.

O. Mercier, Y. Sulai, K. Mackenzie, M. Zannoli, J. Hillis, D. Nowrouzezahrai, and D. Lanman. Fast gaze-contingent optimal decompositions for multifocal displays. *ACM Transactions on Graphics (TOG)*, 36(6):1–15, 2017.

S. Monin, A. C. Sankaranarayanan, and A. Levin. Analyzing phase masks for wide étendue holographic displays. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022a.

S. Monin, A. C. Sankaranarayanan, and A. Levin. Exponentially-wide etendue displays using a tilting cascade. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022b.

J. J. Naji and T. C. Freeman. Perceiving depth order during pursuit eye movement. *Vision research*, 44(26):3025–3034, 2004.

S.-W. Nam, Y. Kim, D. Kim, and Y. Jeong. Depolarized holography with polarization-multiplexing metasurface. *ACM Transactions on Graphics (TOG)*, 42(6):1–16, 2023.

P. Napieralski and F. Rynkiewicz. Modeling human pupil dilation to decouple the pupillary light reflex. *Open Physics*, 17(1):458–467, 2019.

M. Nawrot. Eye movements provide the extra-retinal signal required for the perception of depth from motion parallax. *Vision research*, 43(14):1553–1562, 2003.

R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. *Light field photography with a hand-held plenoptic camera*. PhD thesis, Stanford University, 2005.

N. Padmanaban, Y. Peng, and G. Wetzstein. Holographic near-eye displays based on overlap-add stereograms. *ACM Trans. Graph.*, 38(6), 2019a.

N. Padmanaban, Y. Peng, and G. Wetzstein. Holographic near-eye displays based on overlap-add stereograms. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019b.

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

J.-H. Park. Recent progress in computer-generated holography for three-dimensional scenes. *Journal of Information Display*, 18(1):1–12, 2017.

J.-H. Park. Efficient calculation scheme for high pixel resolution non-hogel-based computer generated hologram from light field. *Optics Express*, 28(5):6663–6683, 2020.

J.-H. Park and M. Askari. Non-hogel-based computer generated hologram from light field using complex field recovery technique from wigner distribution function. *Optics express*, 27(3):2562–2574, 2019.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein. Neural holography with camera-in-the-loop training. *ACM Trans. Graph.*, 39(6):1–14, 2020.

M. Perez-Ortiz and R. K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. *arXiv preprint arXiv:1712.03686*, 2017.

K. Ratnam, R. Konrad, D. Lanman, and M. Zannoli. Retinal image quality in near-eye pupil-steered systems. *Optics Express*, 27(26):38289–38311, 2019.

L. Shi, F.-C. Huang, W. Lopes, W. Matusik, and D. Luebke. Near-eye light field holographic rendering with spherical waves for wide field of view interactive 3d computer graphics. *ACM Trans. Graph.*, 36(6), 2017.

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

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

A. Symeonidou, D. Blinder, A. Munteanu, and P. Schelkens. Computer-generated holograms by multiple wavefront recording plane method with occlusion culling. *Optics express*, 23(17):22149–22161, 2015.

F. Wang, T. Ito, and T. Shimobaba. High-speed rendering pipeline for polygon-based holograms. *Photonics Research*, 11(2):313–328, 2023.

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

G. Westheimer. Directional sensitivity of the retina: 75 years of stiles–crawford effect. *Proceedings of the Royal Society B: Biological Sciences*, 275(1653):2777–2786, 2008.

G. Wetzstein and D. Lanman. Factored displays: improving resolution, dynamic range, color reproduction, and light field characteristics with advanced signal processing. *IEEE Signal Processing Magazine*, 33(5):119–129, 2016.

D. Yang, W. Seo, H. Yu, S. I. Kim, B. Shin, C.-K. Lee, S. Moon, J. An, J.-Y. Hong, G. Sung, et al. Diffraction-engineered holography: Beyond the depth representation limit of holographic displays. *Nature Communications*, 13(1):6012, 2022.

D. Yoo, Y. Jo, S.-W. Nam, C. Chen, and B. Lee. Optimization of computer-generated holograms featuring phase randomness control. *Optics Letters*, 46(19):4769–4772, 2021.

H. Zhang, Y. Zhao, L. Cao, and G. Jin. Fully computed holographic stereogram based algorithm for computer-generated holograms with accurate depth cues. *Optics*

express, 23(4):3901–3913, 2015.
H. Zhang, L. Cao, and G. Jin. Computer-generated hologram with occlusion effect using layer-based processing. *Applied optics*, 56(13), 2017.
Z. Zhang and M. Levoy. Wigner distributions and how they relate to the light field. In *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2009.
F. Zhong, A. Jindal, Ö. Yöntem, P. Hanji, S. Watt, and R. Mantiuk. Reproducing reality with a high-dynamic-range multi-focal stereo display. *ACM Transactions on Graphics*, 40(6):241, 2021.

A APPENDIX

Here, we describe the image formation model and the CGH techniques we use in our setup, including 2.5D, 3D, 4D supervisions. For more comprehensive algorithms, we refer to [Choi et al., 2022, Park, 2017]. All software is implemented in PyTorch [Paszke et al., 2019].

A.1 Image formation model

In a holographic near-eye display, a coherent light source is incident on an SLM with a source field u_{src}. The amplitude or phase of the source field is delayed by a spatially-varying input u_{in}. The manipulated field further propagates, creating a target intensity volume at the desired volume at a distance z off the SLM. We use the angular spectrum method as the free space wave propagation model f with the single sideband encoding [Bryngdahl and Lohmann, 1968, Goodman, 2005]. The resulting complex-valued field u_z is formulated as follows:

$$\begin{aligned} u_z(x,y) &= f\left(u_{\text{SLM}}(x,y), z\right), \\ u_{\text{SLM}}(x,y) &= u_{\text{in}}(x,y)\, u_{\text{src}}(x,y). \end{aligned} \tag{3}$$

$$\begin{aligned} f(u,z) &= \iint \mathcal{F}(u) \cdot \mathcal{H}\left(f_x, f_y, z\right) e^{i2\pi(f_x x + f_y y)} df_x df_y, \\ \mathcal{H}\left(f_x, f_y, z\right) &= \begin{cases} e^{i\left(\frac{2\pi}{\lambda} z \sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}\right)} & \text{if } f_y \geq 0, \\ 0 & \text{if } f_y < 0, \end{cases} \end{aligned} \tag{4}$$

where f_x, f_y denotes the spatial frequency, λ is the wavelength of the light, and $\mathcal{F}$ is the 2D Fourier transform.

A.2 Optimization for binary amplitude SLMs

An SLM modulates the complex-valued field with an input amplitude or phase pattern, and the input is usually quantized into a set of levels Q, (e.g. $\{0,1\}$). Here, we use a 1-bit SLM in amplitude mode, which only supports output of $q_{\text{in}} \in \{0,1\}^{M \times N}$. This SLM can operate at 3600 Hz so the user perceives the time-averaged intensity. In other words, our CGH algorithms aim to obtain the optimal amplitude pattern q_{in} for desired target intensity distributions. Since optimizing binary values is a combinatorial optimization problem which is NP-hard, we relax the binary value q_{in} as an output of quantization function q that takes float value a_{in} as input which we optimize for a specific loss function according to the target data:

$$u_{\text{in}}(x,y) = q_{\text{in}}(x,y) = q\left(a_{\text{in}}(x,y)\right). \tag{5}$$

The quantization process is non-differentiable, which does not allow us to use gradient-descent-based methods. To overcome this, we use the Gumbel-Softmax trick [Jang et al., 2016] for approximating the gradient of the quantization function. Specifically, we update the amplitude values using the following equation:

$$a_{\text{in}}^{(k)} \leftarrow {a_{\text{in}}}^{(k-1)} - \alpha \left(\frac{\partial \mathcal{L}}{\partial q} \cdot \frac{\partial \widehat{q}}{\partial a_{\text{in}}}\right)^T \mathcal{L}\left(s \cdot \left|f\left(a_{\text{in}}^{(k-1)}\right)\right|, a_{\text{target}}\right), \tag{6}$$

where α is the step size, $\mathcal{L}$ is the loss function, q is the quantization function, $\widehat{q}$ is the relaxed quantization function obtained using the Gumbel-Softmax layer, and s is a scaling factor. We present the implementation results comparing different quantization strategies in Supplementary Material.

A.3 2.5D supervision

By leveraging the image formation model and utilizing a gradient descent-based update rule, we can optimize the binary amplitude SLM pattern to accommodate various loss functions as described in [Choi et al., 2022]. First, we produce the 2.5D supervision results in our paper employing the multiplane loss function in Eq. 8. To implement this approach, we first utilize the closest distance matching technique to create a set of binary masks $M^{(k)}$, corresponding to various distances $z^{(k)}$ from the SLM, using the depth map D obtained from an RGB-D input.

$$M^{(k)}(x,y) = \begin{cases} 1, & \text{if } |z^{(k)} - D(x,y)| < |z^{(l)} - D(x,y)|, \forall l \neq k, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Subsequently, we use these binary masks for the multiplane loss, which constrains the wavefront to reconstruct the desired RGB amplitude, denoted as a_{target}, at the relevant distances from the SLM, where $\circ$ represents the element-wise product.

$$\mathcal{L}_{\text{2.5D}} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}\Bigg(M^{(k)} \circ s\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|f\left(q\left(a_{\text{in}}^{(t)}\right), z^{(k)}\right)\right|^2}, M^{(k)} \circ a_{\text{target}}\Bigg). \tag{8}$$

A.4 3D supervision

The 2.5D loss function only restricts the positioning of objects and does not necessarily result in a natural defocus blur for the unconstrained part. To address this, one can assume the amount of defocus occurring at each plane based on the pupil size and penalize all focal slices throughout the volume, ultimately pushing the wavefront toward the desired focal stack using the following loss function:

$$\mathcal{L}_{\text{3D}} = \mathcal{L}\left(s\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|f\left(q\left(a_{\text{in}}^{(t)}\right), z^{\{j\}}\right)\right|^2}, \text{fs}_{\text{target}}\right). \tag{9}$$

The target focal stack can be generated using various techniques, such as RGB-D data, off-the-shelf 3D computer graphics software, or light field data. In our paper, we differentiate between 3D supervision techniques based on how the focal stack is produced. Specifically, we generate the focal stack from RGB-D data, which

we label as 3D w/ RGB-D supervision. In contrast, when the focal stack target is generated from light field data, which offers more realistic occlusion handling, we refer to it as 3D w/ LF supervision.

A.5 4D supervision

It is also possible to obtain an observable light field from the wavefront utilizing the short-time Fourier transform [Padmanaban et al., 2019b, Zhang and Levoy, 2009]. The short-time Fourier transform computes the Fourier transform over a small patch surrounding each pixel, providing information about how each pixel appears from different directions. By exploiting this analytical forward relationship between the observable light field and the wavefront, we can directly penalize the wavefront to create the observable light field, incorporating the short-time Fourier transform into the loss function as presented by [Choi et al., 2022]:

$$\mathcal{L}_{\text{4D}} = \mathcal{L}\left(s\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left|\text{STFT}\left(f\left(q\left(a_{\text{in}}^{(t)}\right),z\right)\right)\right|^2}, \text{lf}_{\text{target}}\right). \tag{10}$$