Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diffraction-engineered holography: Beyond the coherent limits of holographic displays

Daeho Yang
Samsung Electronics
Wontaek Seo
Samsung Advanced Institute of Technology, Samsung Electronics
Hyeonseung Yu
Samsung Electronics
Sun Il Kim
Bongsu Shin
Samsung Advanced Institute of Technology, Samsung Electronics
Chang-Kun Lee
Samsung Advanced Institute of Technology, Samsung Electronics
Seokil Moon
Samsung Electronics
Jungkwuen An
Samsung Electronics https://orcid.org/0000-0003-3918-8402
Jong-Young Hong
Samsung Electronics
Geeyoung Sung
Samsung Advanced Institute of Technology, Samsung Electronics
Hong-Seok Lee (lhs1210@gmail.com)
Samsung Electronics https://orcid.org/0000-0002-3081-7666

Physical Sciences - Article

Keywords:

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1087963/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Diffraction-engineered holography: Beyond the coherent limits of holographic displays

Daeho Yang[1], Wontaek Seo[1], Hyeonseung Yu[1], Sun Il Kim[1], Bongsu Shin[1], Chang-Kun Lee[1], Seokil Moon[1], Jungkwuen An[1], Jong-Young Hong[1], Geeyoung Sung[1], and Hong-Seok Lee[1,*]

[1]Samsung Advanced Institute of Technology, Samsung Electronics, Suwon, Gyeonggi-do, South Korea
[*]lhs1210@samsung.com

November 17, 2021

Holography is considered as one of the most prominent approaches to realize true-to-life reconstructions of objects[1, 2]. However, owing to the limited resolution compared to static holograms[3], state-of-the-art computer generated holograms with dynamic display capabilities reconstruct objects exhibiting various coherent properties, such as interference-induced noise[4–6] and content-dependent defocus blur[7, 8]. Since real world scenes are composed of incoherent light, the coherent properties of reconstructed scenes severely distort the depth perception[9, 10]. Here, we propose a diffraction-engineered hologram, which imitates real world incoherent light by adopting a multi-plane hologram[11, 12], thereby offering a real world-like defocus blur and a photorealistic reconstruction. Our hologram is synthesized by optimizing a wave field to reconstruct numerous varifocal images after propagating the corresponding focal distances where the varifocal images are rendered using a physically-based renderer. By explicitly adopting out-of-focus images as the optimum intensities, the hologram can be synthesized to reconstruct the scene with the correct defocus blur. Moreover, to reduce the computational costs associated with rendering and optimizing, we also demonstrate a network-based synthetic method that requires only an RGB-D image. We experimentally confirm the incoherent-like depth expression of the hologram, while successfully suppressing unnecessary interference in the reconstructed hologram. Our diffraction-engineered hologram offers comparable synthetic time to previously reported methods[13, 14] while presenting significantly more accurate depth cues, moving one step further to reconstructing naturalistic scenes of a virtual world.

Holography is a recording and reconstruction process based on the interference of multiple wave fields[1]. Holograms duplicate the wave field of the recorded object under an appropriate illumination and provide true-to-life reconstructions of three-dimensional(3D) objects[2]. Beyond the reproduction of a recorded object, the computer generated hologram (CGH), which is a numerically calculated hologram of a wave field of non-existing objects, enables the display of arbitrary 3D scenes and provides monocular depth cues, unlike traditional displays[15].

Although holographic displays are free from vergence-accommodation conflict, which causes visual fatigue[16] and a significant reduction in the depth constancy[17], unsolved issues originating from their limited resolution still remain. A real world object scatters light by reflecting light in various directions from the substructures of its rough surface[18], and a static hologram can represent such substructures

with a large effective number of pixels[19]. In contrast, dynamic holograms, of which the resolution is 3 orders of magnitude smaller than that of static holograms[3], cannot spread light without noise because the interference between voxels becomes noticeable as the number of voxels increases[4, 5]. From this perspective, dynamic holograms can be categorized into two different types, namely diffusive holograms and non-diffusive holograms(Fig. 1**a**).

Diffusive holograms spread light up to the maximum diffraction angle bounded by a pixel pitch by introducing high frequency patterns[5, 20–23]. For example, high frequency patterns can be included in holograms by placing voxels with a sufficient separation between them[5, 20], applying random phases[21, 22], and employing point-based methods with physically correct phases[23]. In diffusive holograms, 3D objects can be seen at any position within a viewing angle and out-of-focus objects are blurred as real world objects. However, the image quality is limited by a small number of points or interference between the points[4, 5].

In contrast to diffusive holograms, non-diffusive holograms concentrate on enhancing the image quality of reconstructed scenes. In this case, position-dependent phase offset is imposed in point-based methods to avoid the rapid phase variation of different depth objects[6, 13], phase-retrieval algorithms are adopted to reconstruct single-depth images[24, 25], and quadratic phases are utilized to suppress the speckles[7, 26]. Although non-diffusive holograms tend to exhibit an enhanced image quality, the coherent properties of light become conspicuous due to a reduced numerical aperture and content-dependent defocus blur[7, 8]. Since incoherent light yields a content-independent blur circle diameter, the incorrect diameter destroys the relationship between the depth and the blur, which is crucial in the context of depth perception.[9, 10]. Moreover, the presence of a lucid boundary at the interface between objects with different depths due to interference distorts the perception of the relative depth between objects[27].

Here, we demonstrate a diffraction-engineered hologram (DEH) that presents photorealistic scenes and real world-like defocus blur, breaking the coherent limits of a conventional CGH. For this purpose, we take advantage of the fact that the phase variation of light does not affect the image seen by eyes, but steers the propagating direction of light. As a result, it is possible to find a wave field displaying two or more different images at the same time when the images are displayed at different depths[11, 12, 28, 29]. To obtain such a multi-plane hologram, varifocal images are rendered by a physically-based renderer that properly handles occluded objects and provides an accurate blur circle similar to that of a human eye. The intensity of the wave field of the hologram is optimized to resemble the varifocal images when the propagating distance of the wave field matches the focal distance of each varifocal image. As a result, the DEH achieves both superiorities, namely the image quality of non-diffusive holograms and the depth expression of diffusive holograms. Furthermore, to reduce the computational cost associated with the rendering of varifocal images and the optimization of a complex wave field, we design and train a convolutional neural network. The diffraction-engineered hologram network (DEHNet) synthesizes the complex wave field displaying appropriate blurred images depending on the focal distances while requiring only an RGB-D image as the input. Finally, we confirm the properties of the DEH through simulations and experiments to demonstrate an enhanced depth expression compared to conventional CGHs.

Assuming that a wave field at the $z = 0$ plane is given by $|A(x, y)|e^{i\phi(x,y)}$, the propagated wave field at the $z = d_n$ plane calculated by the angular spectrum method (ASM)[30] is given as

$$\mathrm{Prop}_{d_n}(|A(x,y)|e^{i\phi(x,y)}) = F^{-1}\left\{F\left\{|A(x',y')|e^{i\phi(x',y')}\right\}e^{ik_z d_n}\right\}, \tag{1}$$

where $F(F^{-1})$ is the Fourier (inverse Fourier) transform operator, $e^{ik_z d_n}$ is a propagation kernel with $k_z = \sqrt{k^2 - k_x^2 - k_y^2}$, and $k_x(k_y)$ is the angular wavenumber along the x(y) direction. Here, a notable point of Eq. 1 is the fact that the propagation kernel $e^{ik_z d_n}$ does not alter the amplitude distribution in the Fourier domain, and so the amplitude distribution in the Fourier domain is sustained for every propagation distance. Considering that the diffraction angle is proportional to the spatial frequency[7, 8], the application of a wide frequency range of phases is the only means to achieve sufficient defocus blur

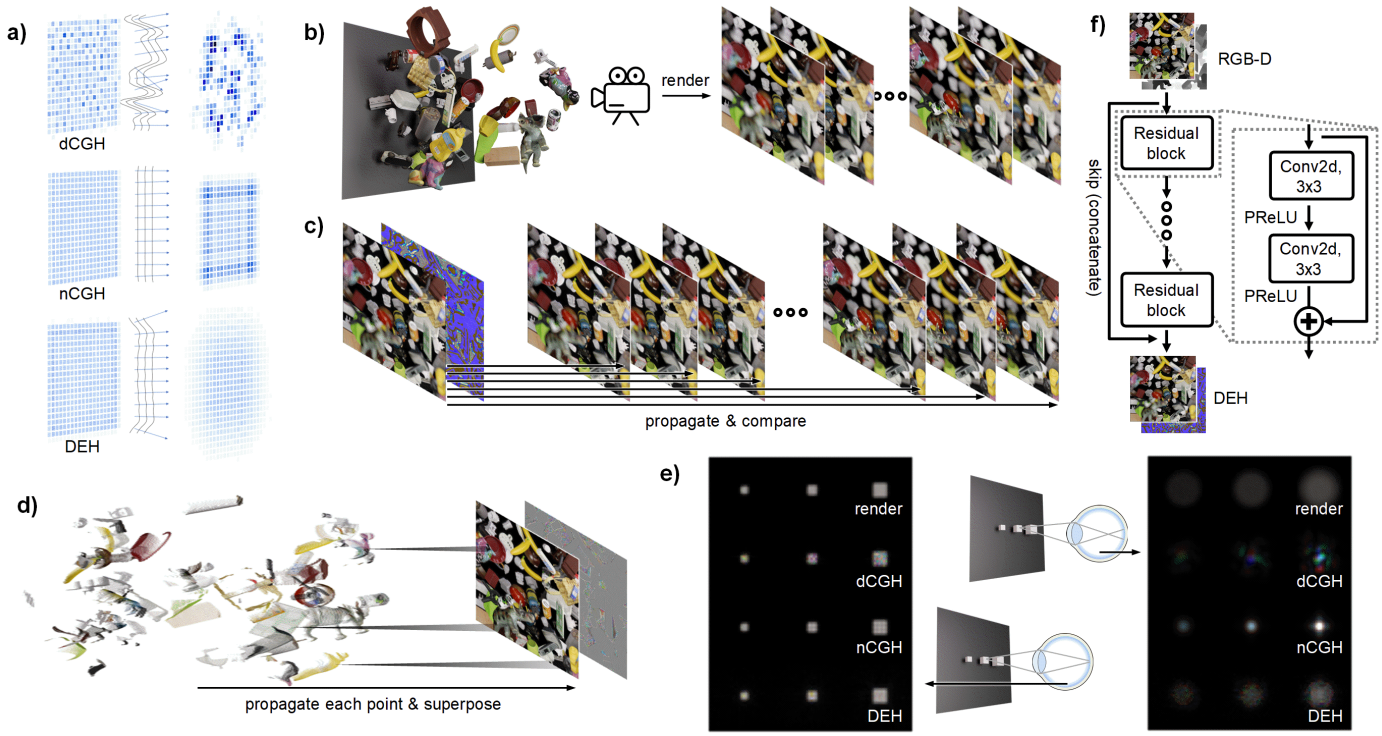


Figure 1: **Schematics of diffraction-engineered holography. a**, The intensity distributions of the diffusive hologram (dCGH), the non-diffusive hologram (nCGH), and the DEH are drawn for two different planes when a blue square is reconstructed at the left side. The black lines represent the phases of the various holograms, while the blue arrows represent the propagating direction of the light. In the DEH, the content-dependent phase at the edge of the square spreads light over a wide angle so defocus blur can be formed at the other focal plane. **b**, Upon varying the focal distance of the camera, varifocal images are rendered using a physically-based renderer. **c**, The wave field is optimized to satisfy all varifocal images at each depth in the DEH. **d**, The nCGH synthesizes a hologram by propagating each voxel and superposing the propagated voxels. **e**, In-focus and out-of-focus intensities of the rendered case, the dCGH, the nCGH, and the DEH are simulated (from top to bottom). The side lengths of the cubes are 3, 4, and 6 (from left to right). **f**, The convolutional neural network synthesizes a wave field from an all-in-focus image and an all-in-focus depth map.

unless the intensity itself is composed of wide range of frequencies.

However, the majority of high quality non-diffusive CGH (nCGH) algorithms fix the phase as zero or as a position-dependent formula[6, 13, 30] to avoid speckles, thereby leaving the content-dependent defocus blur unsolved. The DEH starts from this point. DEH is calculated by optimizing a wave field to possess a content-dependent phase so that the propagated wave field forms a clear image at the object-existing plane while forming a blurred image at other planes. As a target image for each propagated distance, we used varifocal images generated by a rendering process with changing focal distance of a camera to ensure that occlusion-considered blur is efficiently reflected(Fig. 1**b**). After simulating the propagated intensity of the wave field using the ASM, we calculated the mean square error (MSE) between the intensity and the varifocal image of which the focal distance is equal to the propagation distance(Fig. 1**c**). The wave field is compared with tens of varifocal images and it is updated using a gradient descent method. The optimization is iterated until the change of the wave field is negligible and the wave field is subsequently propagated by the average focal distance of the varifocal images.

Compared to other researches[13, 14] employing learning-based methods or optimization methods, occlusions and defocus blur can be reflected on the reconstructed scene by means of explicitly comparing the propagated intensities and defocused images. Furthermore, to reconstruct sharply focused objects, the wave field is also compared with an all-in-focus image when the propagation distance is close to the depth of the objects (see the Methods for further details). Standard phase retrieval algorithms, *e.g.* the

iterative Fourier transform algorithm, can be used in multi-plane holograms[11, 28, 29], but gradient descent optimization is employed to compare the wave field with the depth-weighted all-in-focus image.

In summary, the total loss function $\mathcal{L}$ for optimization is given by,

$$\begin{aligned}\mathcal{L} = \sum_{n=1}^{N} \Bigg[& \left\langle \left| |\mathrm{Prop}_{d_n}(|A(x,y)|e^{i\phi(x,y)})|^2 - I_{\mathrm{d_n}} \right|^2 \right\rangle \\ & + \beta \left\langle \left| \left(|\mathrm{Prop}_{d_n}(|A(x,y)|e^{i\phi(x,y)})|^2 - I_{\mathrm{AIF}} \right) e^{-\left(\gamma \frac{\mathrm{D}_n - d_n}{d_0 - d_N}\right)^2} \right|^2 \right\rangle \Bigg],\end{aligned} \tag{2}$$

where N is the number of varifocal images, $I_{\mathrm{d_n}}$ is the intensity of the varifocal image at a focal distance d_n, I_{AIF} is the intensity of the all-in-focus image, D_n is the depth map normalized from d_0 to d_N with a focal distance d_n, β is the user-defined loss weight, and γ is the user-defined depth attention weight. Here, the depth map with defocus blur depending on the focal distance is used instead of an all-in-focus depth map to reflect the occlusion (see the Methods for further details). The first term in Eq. 2 represents the MSE of the propagated wave field compared to the varifocal images, while the second term represents the MSE of the propagated wave field compared to the depth-weighted all-in-focus image. In contrast to a DEH, conventional methods[13, 30] construct holograms by propagating each 3D point for a particular distance depending on its depth value and superposing the propagated voxels(Fig. 1**d**). The method simulates propagation of the points by the ASM and also handles occlusion by ignoring the backside wavefront when the backside and frontside wavefronts overlap.

Holograms depicting a scene with different-sized cubes were synthesized and in-focus (out-of-focus) conditions of the holograms were simulated as shown in Fig. 1**e**. Even in the out-of-focus conditions, the defocus blur of the nCGH cannot be seen, especially for the large cube, due to the content-dependent defocus blur[7, 8]. Coherent propagation of the wave field forms a Fresnel diffraction pattern which differs from the defocus blur of incoherent light so the depth perception can be distorted[9]. In contrast, the out-of-focus image of the DEH displays clear defocus blur even if the diameter of the blur circle is slightly smaller than that of the rendered image. However, the most significant drawback of the DEH is its computation power, since a number of varifocal images are required in addition to an optimization procedure. Since nCGH can be synthesized using only RGB-D images, DEHs are not practical in the majority of real-time applications.

To overcome such issues, a neural network (DEHNet) is trained to obtain a DEH from RGB-D images(Fig. 1**f**). The network is composed of 34 convolutions with 12 channels except for the last layer which includes a concatenated shortcut. Non-linearity and a wide receptive field are more important than hidden features so the number of channels are selected as small as possible to increase the number of convolutions and activations under a restricted computation resource. The training dataset consists of 3000 different scenes and each of these scenes contains 21 varifocal images, an all-in-focus color image, 21 varifocal depth maps, and an all-in-focus depth map. After training, the DEHNet can synthesize an optimal wave field which can reconstruct appropriate blurred and sharply focused images while considering occlusions, and this can be achieved using only an all-in-focus color image and an all-in-focus depth map.

Figure 2 shows the simulated results for the DEH, the DEHNet, and the nCGH when the focus is adjusted to the frontside or backside of the scene. The diffusive hologram is excluded from the comparison because its image quality is not compatible with other methods unless other techniques, such as the time-multiplexing technique, are adopted simultaneously. One of the differences between the nCGH and the DEH is the vivid boundary at the interface of the objects which are located at different depths as shown in the enlarged image of Fig. 2. An abrupt phase variation at the interface leads to two coherent beams with different phases coinciding at the interface; the constructive and destructive interferences then build a sharp boundary. Since blurred and sharply focused edges at the occlusion boundary are used to judge the relative depths between objects[27], the presence of a distorted blur at an edge can be considered one of the most serious defects. Moreover, when a hole exists in an object, the hole is distorted by the depth difference between the object and the background.

Figure 2: **Simulation results for the DEH, the DEHNet, and the nCGH.** With the exception of the all-in-focus image (top-left) and the depth map (bottom-left), the top images correspond to the front focus images and the bottom images correspond to the rear focus images. The PSNR values (in dB) and the SSIM values are marked on the top right corner of each image. The smaller images represent enlarged views of the larger images. The ASM was used to simulate different focal planes.

The perceptual image quality, including defocus blur as well as speckle noise, can be measured quantitatively by evaluating the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) compared to the rendered images. While the optimized DEH exhibits the best PSNR (26.1) and SSIM (0.88) values, the DEHNet also gives compatible results. In contrast, the nCGH gives significantly lower PSNR (19.8) and SSIM (0.67) values. Here, the second term in Eq. 2 boosts the image quality of the in-focus objects, which results in a slight reduction in the PSNR. Without the second term, the PSNR increases slightly (0.6∼0.8) although the image quality at the focal plane is reduced.

In order to concretely validate the DEHNet, an experimental demonstration is necessary. In an optical reconstruction, an amplitude-only spatial light modulator (SLM) with a 1920×1080 (FHD) resolution is used instead of a complex SLM. It is well known that an amplitude SLM can be used as a complex SLM by means of spatial filtering, although the spatial bandwidth of the SLM is lost [31]. As confirmed by the simulation, the defocus blur is much weaker and a vivid boundary exists near the interface of the different-depth objects in the nCGH. As a consequence, it is difficult to perceive the depth of the 3D scene in the nCGH. This tendency is more apparent in the enlarged images shown in Fig. 3. Details regarding the experimental setup and parameters can be found in the Methods section.

Figure 4 shows the inference times of the various CGH-generation methods, which were evaluated on an NVIDIA V100 GPU using the FHD resolution images. Since an optimization-based DEH requires 500 iterations, the method requires more than 1 minute to synthesize a hologram with a superior image quality. However, we achieved a frame rate of 62 Hz using a quantized network, of which the weights were quantized to 8-bit integers, while losing only ∼0.5 dB of the PSNR compared to the optimization method. Although it was not mentioned before, all simulation results and experimental results for the DEHNet were achieved using the quantized network.

Figure 3: **Experimental results of the DEHNet and the nCGH.** The top images correspond to the front focus images and the bottom images correspond to the rear focus images. The small images represent enlargements of the corresponding holograms, as indicated by the white squares. The bright spots close to the center can be attributed to the low performance of the anti-reflection coating of the objective lens.

The PSNR and SSIM were evaluated for two datasets with different resolutions to quantitatively measure the image quality. One dataset is composed of 512×512 resolution images (Fig. 4**b**) as in the case of the training dataset, while the other dataset is composed of FHD resolution images (Fig. 4**c**). The indicated metrics represent the mean values of the comparison results between all varifocal images and the corresponding holograms so the smoothness of defocus blur and the sharpness of the focused object are both reflected. In the 512 (FHD) resolution dataset, the DEHNet provides a 6.5 (6.3) dB enhancement in the PSNR and a 0.15 (0.07) enhancement in the SSIM compared to the nCGH. Both of the evaluation datasets are rendered with textures that differ from that of the training dataset to ensure that the performance of the trained network is not restricted to the training dataset.

When the holograms are synthesized using real world images instead of rendered images, it should be pointed out that incorrect values from the captured depth maps can induce severe noise. In the majority of cases, real world-captured depth maps include depth holes and incorrect depth values[32] so the interference pattern distorts the objects when the object boundaries of the depth map are not consistent with those of the RGB image (Extended Fig. 1 and Extended Fig. 2). In contrast to the nCGH producing interference-induced black lines at the boundaries of noisy depth, the DEH provides noise-suppressed images at these boundaries. In some applications using measured depth maps, *e.g.* video see-through displays, the DEH would therefore give a superior image quality than the nCGH.

In summary, we proposed and experimentally confirmed that the DEH depicts an arbitrary 3D scene with a full range of depth without distortion of the depth perception. At the same time, we demonstrated that the network can convert RGB-D images into DEHs in real time and that the increments in the PSNR and SSIM metrics are substantial. We expect the DEHs could be widely used in holographic displays for virtual and augmented realities offering real world-like 3D displays using currently available display devices.

It should also be noted here that in some works, the multi-plane hologram refers to the hologram reconstructing multiple objects at different depths, as an antonym of the hologram that reconstructs multiple objects at a single depth[14]. In contrast, we use the term to represent a hologram that can reconstruct numerous full-size images at the same time depending on the focal distance. As the latter hologram, our experiment shows a greatly enhanced image quality in comparison with those reported

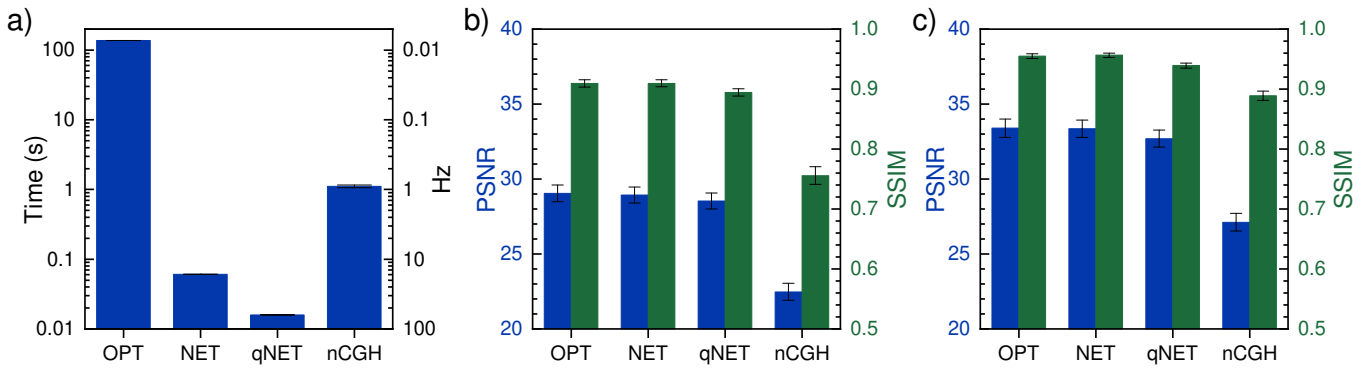


Figure 4: **Performance comparison. a**, Inference times. OPT refers to the DEH calculated by optimization, NET refers to the DEH calculated by the neural network, and qNET refers to the DEH calculated by the quantized neural network. We achieved a frame rate of 62 Hz in qNET, which is 8600 (70) times faster than that obtained in OPT (nCGH). The OPT inference time does not include the rendering time of the varifocal images. **b**, The PSNR and SSIM were evaluated for the 512 resolution dataset. The presented PSNR and SSIM values represent the mean values of all images in the dataset with respect to the 21 rendered images for each scene. OPT, NET, and qNET have similar PSNR values (29.0, 28.9, and 28.5 dB, respectively) and SSIM values (0.909, 0.910, and 0.894, respectively), while nCGH gives significantly lower PSNR (22.5 dB) and SSIM (0.756) values. **c**, The PSNR and SSIM were evaluated for the FHD resolution dataset. For the OPT, NET, qNET, and nCGH methods, the SSIM (PSNR) values were given by 0.955 (33.4 dB), 0.957 (33.4 dB), 0.939 (32.7 dB), and 0.889 (27.1 dB), respectively. The error bars represent the standard deviations between scenes.

previously[11, 12, 28, 33] despite the fact that more than 20 images were used as target images. The degraded image quality in previous experiments mainly originated from the high frequency patterns that almost reached the pixel-pitch-limited frequency, since a phase-only SLM or an amplitude-only SLM was used instead of a complex SLM[11, 29]. Our experiment confirms that it is possible to reconstruct multiple intensities with great fidelity when the target intensities are gradually varied, suggesting the feasibility of real-time applications of multi-plane holograms, such as holographic optical tweezers[33], one-step volumetric printings[34], and volumetric displays[35].

methods

Determining the diameter of the blur circle

To construct large field of view (FoV) display systems, an SLM is magnified by a lens array. As a consequence, the maximum propagation distance of the hologram that allows the reconstruction of a virtual image with a depth from d to infinity is determined by the effective focal length of the lens array. By approximating the lens array as a thin lens, the maximum propagation distance of the hologram, Δz, can be calculated as[36],

$$\Delta z \approx \frac{1}{d}\frac{\Delta x^2 \mathrm{res}^2}{4\tan^2(\mathrm{FoV}/2)} \tag{3}$$

where Δx is the pixel pitch of the SLM, res is a resolution of the display, FoV is the field of view of the system, d is the virtual image distance of the floating object synthesized by the hologram, and the virtual image distance of the display is set to infinity. If we consider a 55° FoV, a 4K resolution, a 7.2 um pixel pitch, and d=0.35 m, then Δz=2 mm is obtained from Eq. 3.

Under the specific display parameters that were considered herein, it is possible to calculate the diameter of a blur circle of a human eye when the eye is focused on infinity while the object synthesized by the hologram is floating at a distance of d. The diameter of a blur circle of an eye in units of display

pixels, $\mathrm{CoC_{eye}}$, is given by,

$$\mathrm{CoC_{eye}} = \frac{A \cdot \mathrm{res}}{2d \cdot \tan(\mathrm{FoV}/2)} \tag{4}$$

where A is the pupil diameter. If the wave field of the hologram is partially blocked by the iris, the image quality degrades by the noise of the blocked wave field. Considering that the diameter of a pupil is larger than 1.5 mm in the majority of cases [37, 38], A is set to 1.5 mm to avoid image degradation originating from a partially blocked wave field. From the above parameters, $\mathrm{CoC_{eye}}$ is 15 pixels and the aperture size of the rendering camera is set to satisfy the diameter of a blur circle of the rendered images.

Although a diameter of defocus blur of an nCGH can be enlarged by increasing the propagation distance, achieving a blur circle equivalent to that of a human eye is only possible under a small FoV ($\sim$10 °). For example, if we increase the propagation distance to enlarge the diameter of the blur circle, the virtual image distance of the object(d) comes closer and the blur circle diameter of the eye($\mathrm{CoC_{eye}}$) is also increased. As a result, an increase in the diameter of the defocus blur under a fixed propagation distance is required to attain a human eye-equivalent defocus blur with a holographic display.

Experimental details

In the experiment, an FHD resolution amplitude-only LCoS (liquid crystal on silicon) with a pixel pitch of 7.2 μm was employed (Extended Fig. 3), and the distance between the minimum and maximum depths was set to 2 mm. The dispersion diameters by the pixel pitch diffraction are 25 (red), 20 (green), and 18 pixels (blue) under 2 mm light propagation. Considering that the maximum diameter of defocus blur of the rendered images is 15 pixels, the propagation distance should be longer than 1.7 mm. Since the modulated intensity non-linearly depends on the assigned values of the pixels, the amplitude was calibrated by measuring output values for each input pixel value. An off-axis hologram was adopted and the grating period was set to 0.25 of its maximum period to avoid unwanted noise. The Burch encoding method[39] was used to project the complex wave field onto real values. With an adjustable 2D slit, zeroth order and higher order diffractions are blocked. As a light source, laser diodes with wavelengths of 638, 515, and 460 nm were used and were sequentially illuminated on the LCoS. To remove speckles caused by the coherence of the lasers, the holographic diffuser was rotated at the focused spot of the laser beams.

Phase noise of the amplitude-only SLM

Due to the properties of liquid crystals, it is inevitable that the amplitude-only SLM modulates the phase. The noise from such phase modulation can be avoided if an appropriate grating phase is applied. Assuming that an amplitude modulation is given by $f(x)$ and unwanted phase modulation is given by $\exp\{ip_1 f(x) + ip_2 f(x)^2\}$, then the wave field at the SLM is given as $f(x)e^{ip_1 f(x)+ip_2 f(x)^2}$. Here, we approximated the unwanted phase modulation as a second order polynomial function of the amplitude modulation. To expand the expression, we employed the Jacobi-Angler expansion, $e^{ikz\cos\theta} = \sum_{n=-\infty}^{\infty} i^n J_n(z) e^{in\theta}$, where $J_n(z)$ is the n-th Bessel function of the first kind. Using a Fourier series expansion, $f(x) = \sum_k F_k \cos(kx + \phi_k)$, the wave field at the SLM can be expressed as,

$$\begin{aligned} f(x)e^{ip_1 f(x)+ip_2 f(x)^2} &= f(x)e^{ip_1(\sum_k F_k\cos(kx+\phi_k))+ip_2(\sum_k F_k\cos(kx+\phi_k))^2} \\ &= f(x)\prod_k\sum_n i^n J_n(p_1F_k)e^{in(kx+\phi_k)} \\ &\times \prod_{k,l}\sum_n i^n J_n(p_2F_kF_l/2)e^{in((k+l)x+\phi_k+\phi_l)} \\ &\times \prod_{k,l}\sum_n i^n J_n(p_2F_kF_l/2)e^{in((k-l)x+\phi_k-\phi_l)}. \end{aligned} \tag{5}$$

Fortunately, p_1, p_2, and F_k are less than 1 in our experiment, and so $J_n(z)$ with $|n| \ll 1$ can be neglected for those cases. As a result, Eq. 5 can be approximated as,

$$
\begin{aligned}
& f(x)e^{ip_1 f(x)+ip_2 f(x)^2} \\
& \approx f(x)\prod_k J_0(p_1F_k)\left(\prod_{k,l} J_0(p_2F_kF_l/2)\right)^2 \times \left[\sum_m \frac{iJ_1(p_1F_m)}{J_0(p_1F_m)}e^{i(mx+\phi_m)}\right. \\
& \left. + \sum_{m,n}\frac{2iJ_1(p_2F_mF_n/2)}{J_0(p_2F_mF_n/2)}\left(e^{i((m+n)x+\phi_m+\phi_n)}+e^{i((m-n)x+\phi_m-\phi_n)}\right)+\mathcal{O}((p_1F_k)^2)+\mathcal{O}((p_2F_k^2)^2)\right].
\end{aligned} \tag{6}
$$

As we can see from Eq. 6, if a grating phase with a period $e^{ik_{\text{prism}}x}$ is applied, then $e^{ik_{\text{prism}}x}$, $e^{-ik_{\text{prism}}x}$, $e^{2ik_{\text{prism}}x}$, and a constant term are generated. Moreover, Burch encoding[39] generates its conjugate term $e^{-ik_{\text{prism}}x}$ and its phase noise-induced terms. As a result, the $e^{ik_{\text{prism}}x}$, $e^{-ik_{\text{prism}}x}$, $e^{-ik_{\text{pitch}}x+2ik_{\text{prism}}x}$, $e^{-ik_{\text{prism}}x}$, $e^{ik_{\text{prism}}x}$, $e^{ik_{\text{pitch}}x-2ik_{\text{prism}}x}$ terms exist, where k_{pitch} is the wavenumber of the SLM pixel pitch and the terms such as $e^{ik_{\text{pitch}}x-2ik_{\text{prism}}x}$ are created by the black matrix of the SLM. When the frequency of the grating phase is one third of the spatial frequency of the pixel pitch, our signal term $e^{ik_{\text{prism}}x}$ overlaps with the noise term $e^{ik_{\text{pitch}}x-2ik_{\text{prism}}x}$ and the noise cannot be filtered. To avoid such noise, the frequency of the grating phase was set to one quarter or less of the spatial frequency of the pixel pitch.

Generation of the training dataset

The objects in the 3D scene were randomly sampled from publicly available datasets[40–43] and each scene was rendered by Blender to have 21 varifocal images[44]. The textures of the objects used in the training stage were randomly sampled from the CC0 texture library and the textures of the objects used in the evaluation stage were sampled from the "Benchmark for 6D Object Pose Estimation" datasets[40–43]. The colors, orientations, and intensities of the light sources were randomly sampled while the maximum intensity was restricted to prevent overexposure. When a scene is overexposed, intensity sums of each varifocal image could be different because the intensities become clipped. Since the propagation of light conserves its total energy, varifocal images with inconsistent intensity sums cannot be constructed with a single wave field.

The focal planes of each scene were equally spaced while the distances between the camera and the objects were significantly longer than the distances between the different objects to symmetrically blur either side of the focal plane. The symmetric blur in the rendered images is consistent with the asymmetric blur of an eye when a tiny display is magnified and projected to the eye. With the exception of the background, the pixel-wise statistics of the depth distribution were made almost uniform to prevent overfitting to a particular depth during training.

Parameters of the loss function and the depth map with defocus blur

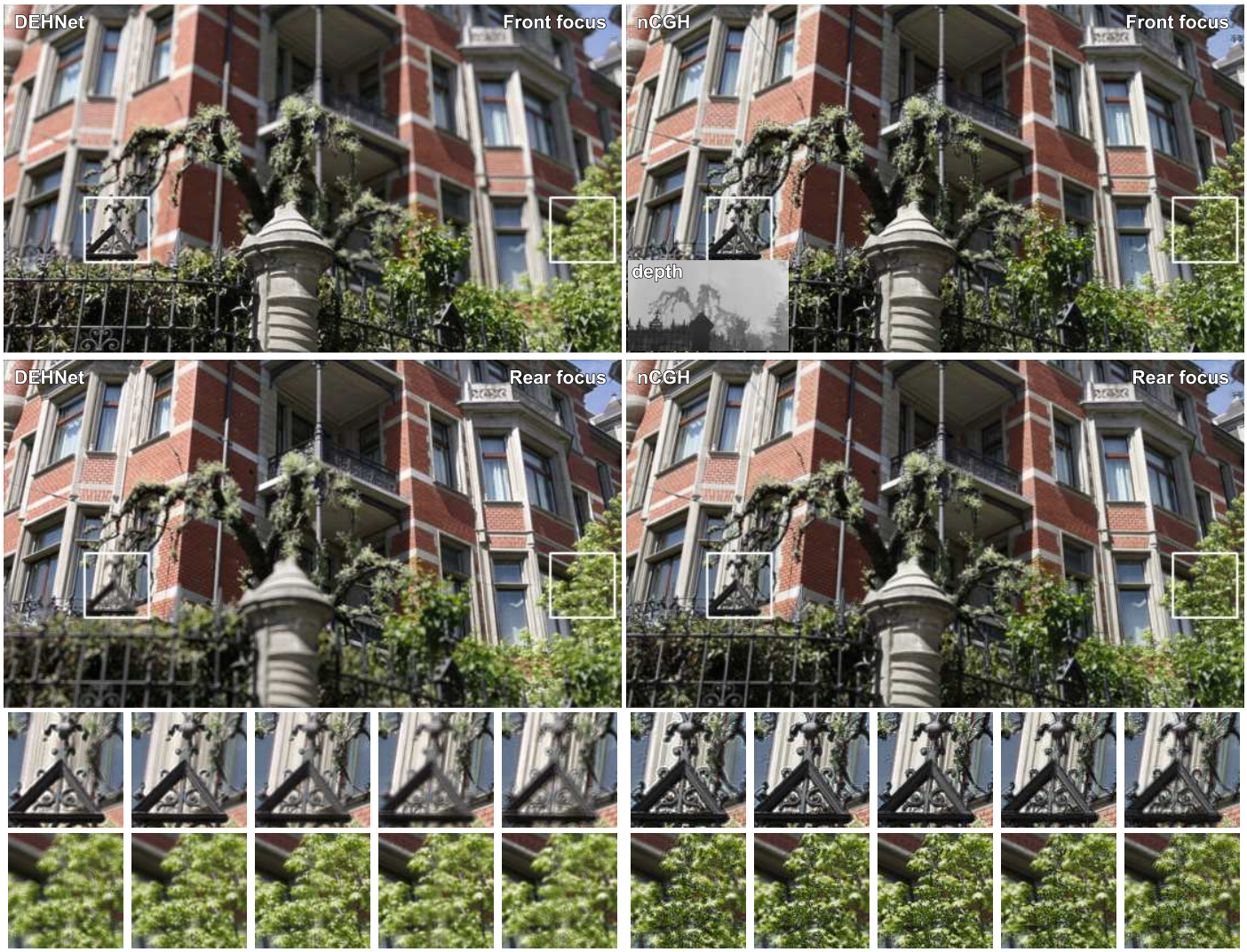
Since the objects in the scene can have any depth, the number of varifocal images was selected to be 21 pixels larger than the maximum diameter of the blur circle, while γ was fixed to 40 to avoid the simultaneous focusing of an object at two different focal planes. For an arbitrary object, the number of out-of-focus images (20) is significantly larger than the number of in-focus images (1) and so the reconstructed scene of the DEH is more influenced by the blurred images than the focused image. Thus, to apply a similar or higher weight to an in-focus image of objects, β was set to 20.

Although defocus blur is not applied to a depth map in the majority of applications, we used a defocus-blurred depth map during the optimization and training processes to consider occlusion. If we assume that one object is located at the front of the scene and another object is located at the rear of the scene, a blur circle of the rear object does not invade a focused image of the front object when the front object is focused. In contrast, a blur circle of the front object invades a focused image of the rear

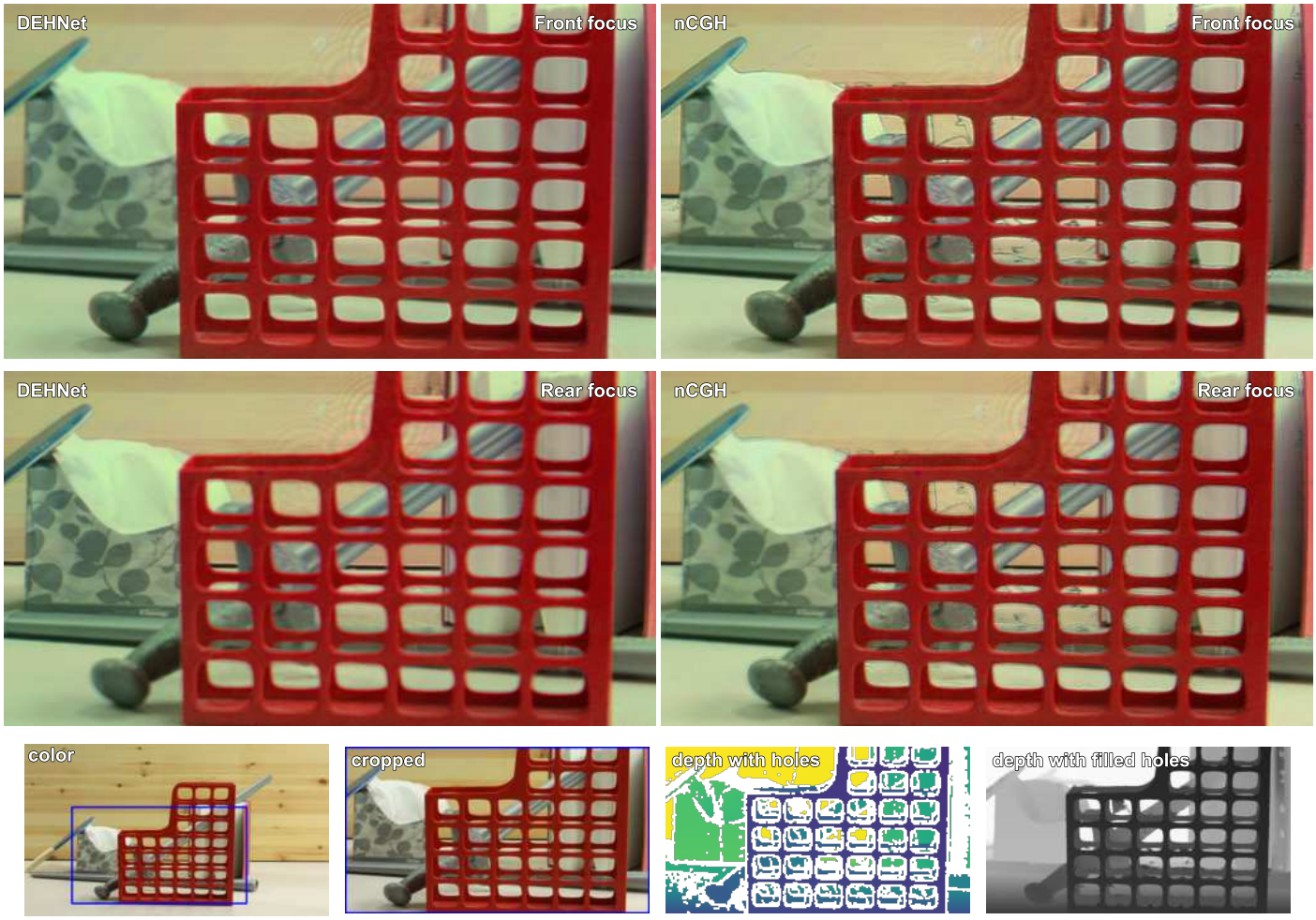
object when the rear object is focused (Extended Fig. 4). Assuming that an all-in-focus depth map is used when comparing the depth-weighted all-in-focus image and the intensity of the hologram for the rear plane of focus (second term of Eq. 2), the pixel weights of the rear object close to the front object are high even if the blur circle degrades the image quality. As a result, the loss function has a lower value when a sharply focused image is reconstructed near the boundary of the front object, ignoring the defocus blur of the front object. Such circumstances can be avoided when the defocus-blurred depth map is used for the second term of Eq. 2 since the rear object occupies a smaller area in this depth map than in the all-in-focus depth map for the rear plane of focus.

Training of the neural network

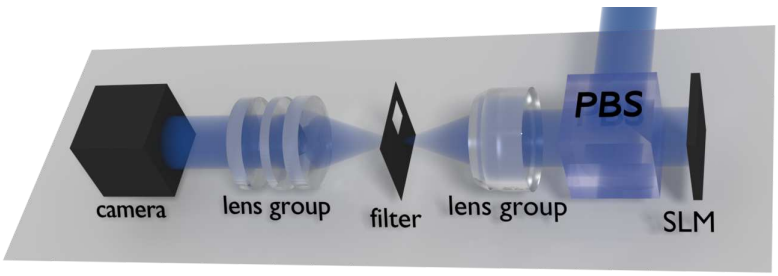
In the first stage of training, we used batch normalization layers in front of activation layers. When the validation loss stopped decreasing, the batch normalization layers and convolution layers were manually fused using running means and running variances. After fusing the batch normalization layers and convolution layers, the fused layers were trained again with the same dataset until the validation loss stopped decreasing. The training process took approximately 60 h using an NVIDIA V100 GPU. The trained neural network was symmetrically quantized using the TensorRT library and the same training dataset was fed to calibrate the quantization parameters.



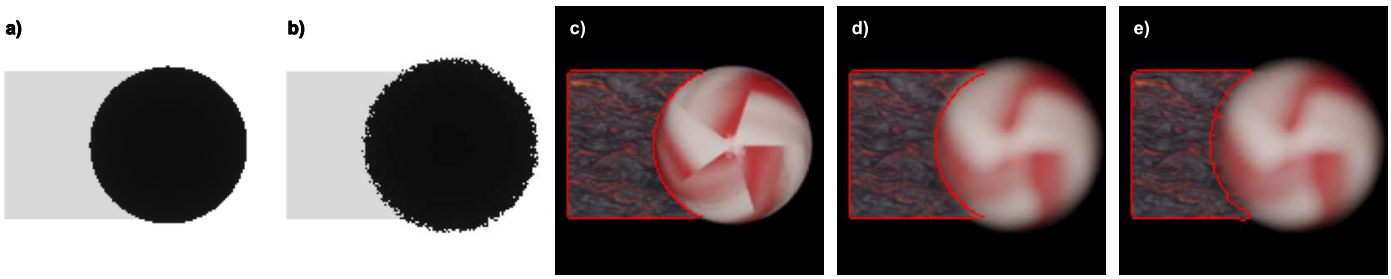
Extended Figure 1: **Simulation results obtained using a real world RGB-D image.** Using a real world-captured image[45], a DEH and an nCGH were synthesized and their intensities were simulated wherein the total image size was resized to 1280 × 720. The large images show the front- and rear-focused images, while the small images show the focus-dependent images (from left to right, 0, 0.6, 1.2, 1.8, 2.4, and 3.0 diopter). The first row of small images shows the front objects and the rear objects simultaneously, while the second row of small images shows the noise on the leaves originated from the imperfect depth map.



Extended Figure 2: **Experimental results obtained using a real world RGB-D image.** Using a real world-captured RGB-D image[32], a DEH and an nCGH were synthesized and optically reconstructed. Since the real world depth map includes depth holes, a monocular depth estimation algorithm[46] was adopted to fill the holes. The optically reconstructed images are cropped to show the details. Black lines caused by wave interference can be seen in the nCGH results but not in the DEH results. Moreover, defocus blur can be perceived only in the DEH results. "Color" represents the all-in-focus color image, "cropped" represents the cropped image, "depth with holes" represents the measured depth map depicting the depth holes with a white color, and "depth with filled holes" represents the hole-filled depth map.



Extended Figure 3: **Schematic representation of the experimental setup.** Collimated RGB lasers were illuminated on an SLM through a polarizing beam splitter (PBS). After the Fourier plane is formed by the first lens array, zeroth order and higher orders of the grating phase diffraction implemented on the SLM were blocked by the filter. The reconstructed hologram was captured by a camera with a second lens array.



Extended Figure 4: **All-in-focus depth map and defocus-blurred depth map overlaid on a color image. a**, The all-in-focus depth map. **b**, The defocus-blurred depth map for the rear plane of focus. While acquiring the depth values in the area of the defocus blur, the depth values of the front object and that of the rear object were randomly sampled by the renderer. To ignore the depth values of the rear object near the boundary, the depth maps were acquired multiple times and the most front values among the numerous depth maps were used. **c**, The boundary of the rear object of the all-in-focus depth map is marked as a red line in the all-in-focus image. **d**, The boundary of the rear object of the all-in-focus depth map is marked as a red line in the rear-focus image. **e**, The boundary of the rear object of the rear-focus depth map is marked as a red line in the rear-focus image.

References

[1] Gabor, D. A new microscopic principle. *Nature* **161**, 777–778 (1948).

[2] Yaraş, F., Kang, H. & Onural, L. State of the art in holographic displays: a survey. *Journal of Display Technology* **6**, 443–454 (2010).

[3] Pan, Y., Liu, J., Li, X. & Wang, Y. A review of dynamic holographic three-dimensional display: algorithms, devices, and systems. *IEEE Transactions on Industrial Informatics* **12**, 1599–1610 (2015).

[4] Haist, T. & Osten, W. Holography using pixelated spatial light modulators—part 1: theory and basic considerations. *Journal of Micro/Nanolithography, MEMS, and MOEMS* **14**, 041310 (2015).

[5] Makowski, M. Minimized speckle noise in lens-less holographic projection by pixel separation. *Optics Express* **21**, 29205–29216 (2013).

[6] Maimone, A., Georgiou, A. & Kollin, J. S. Holographic near-eye displays for virtual and augmented reality. *ACM Transactions on Graphics (Tog)* **36**, 1–16 (2017).

[7] Shimobaba, T. & Ito, T. Random phase-free computer-generated hologram. *Optics Express* **23**, 9549–9554 (2015).

[8] Ko, S.-B. & Park, J.-H. Speckle reduction using angular spectrum interleaving for triangular mesh based computer generated hologram. *Optics Express* **25**, 29788–29797 (2017).

[9] Mather, G. & Smith, D. R. Blur discrimination and its relation to blur-mediated depth perception. *Perception* **31**, 1211–1219 (2002).

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

[11] Makowski, M., Sypek, M., Kolodziejczyk, A., Mikula, G. & Suszek, J. Iterative design of multiplane holograms: experiments and applications. *Optical Engineering* **46**, 045802 (2007).

[12] Makey, G. *et al.* Breaking crosstalk limits to dynamic holography using orthogonality of high-dimensional random vectors. *Nature Photonics* **13**, 251–256 (2019).

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

[14] Peng, Y., Choi, S., Padmanaban, N. & Wetzstein, G. Neural holography with camera-in-the-loop training. *ACM Transactions on Graphics (TOG)* **39**, 1–14 (2020).

[15] Geng, J. Three-dimensional display technologies. *Advances in Optics and Photonics* **5**, 456–535 (2013).

[16] 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**, 33–33 (2008).

[17] Watt, S. J., Akeley, K., Ernst, M. O. & Banks, M. S. Focus cues affect perceived depth. *Journal of Vision* **5**, 7–7 (2005).

[18] Warnick, K. F. & Chew, W. C. Numerical simulation methods for rough surface scattering. *Waves in Random Media* **11**, R1 (2001).

[19] Colburn, W. & Haines, K. Volume hologram formation in photopolymer materials. *Applied Optics* **10**, 1636–1641 (1971).

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

[21] Zhao, Y., Cao, L., Zhang, H., Kong, D. & Jin, G. Accurate calculation of computer-generated holograms using angular-spectrum layer-oriented method. *Optics Express* **23**, 25440–25449 (2015).

[22] Zhao, T., Liu, J., Duan, J., Li, X. & Wang, Y. Image quality enhancement via gradient-limited random phase addition in holographic display. *Optics Communications* **442**, 84–89 (2019).

[23] Tsang, P., Poon, T.-C. & Wu, Y. Review of fast methods for point-based computer-generated holography. *Photonics Research* **6**, 837–846 (2018).

[24] Chang, C. *et al.* Speckle-suppressed phase-only holographic three-dimensional display based on double-constraint gerchberg–saxton algorithm. *Applied Optics* **54**, 6994–7001 (2015).

[25] Chakravarthula, P., Peng, Y., Kollin, J., Fuchs, H. & Heide, F. Wirtinger holography for near-eye displays. *ACM Transactions on Graphics (TOG)* **38**, 1–13 (2019).

[26] Pang, H., Wang, J., Cao, A. & Deng, Q. High-accuracy method for holographic image projection with suppressed speckle noise. *Optics Express* **24**, 22766–22776 (2016).

[27] Marshall, J. A., Burbeck, C. A., Ariely, D., Rolland, J. P. & Martin, K. E. Occlusion edge blur: a cue to relative visual depth. *JOSA A* **13**, 681–688 (1996).

[28] Dorsch, R. G., Lohmann, A. W. & Sinzinger, S. Fresnel ping-pong algorithm for two-plane computer-generated hologram display. *Applied Optics* **33**, 869–875 (1994).

[29] Makowski, M., Sypek, M., Kolodziejczyk, A. & Mikula, G. Three-plane phase-only computer hologram generated with iterative fresnel algorithm. *Optical Engineering* **44**, 125805 (2005).

[30] Matsushima, K. & Shimobaba, T. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. *Optics Express* **17**, 19662–19673 (2009).

[31] Arrizón, V., Méndez, G. & Sánchez-de La-Llave, D. Accurate encoding of arbitrary complex fields with amplitude-only liquid crystal spatial light modulators. *Optics Express* **13**, 7913–7927 (2005).

[32] Scharstein, D. *et al.* High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, 31–42 (Springer, 2014).

[33] Sinclair, G. *et al.* Interactive application in holographic optical tweezers of a multi-plane gerchberg-saxton algorithm for three-dimensional light shaping. *Optics Express* **12**, 1665–1670 (2004).

[34] Shusteff, M. *et al.* One-step volumetric additive manufacturing of complex polymer structures. *Science Advances* **3**, eaao5496 (2017).

[35] Smalley, D. *et al.* A photophoretic-trap volumetric display. *Nature* **553**, 486–490 (2018).

[36] Saleh, B. E. & Teich, M. C. *Fundamentals of Photonics* (John Wiley & Sons, 2019).

[37] Alexandridis, E. Pupil size. In *The Pupil*, 11–12 (Springer, 1985).

[38] Ren, P. *et al.* Off-line and on-line stress detection through processing of the pupil diameter signal. *Annals of Biomedical Engineering* **42**, 162–176 (2014).

[39] Burch, J. A computer algorithm for the synthesis of spatial frequency filters. *Proceedings of the IEEE* **55**, 599–601 (1967).

[40] Hodan, T. *et al.* Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 19–34 (2018).

[41] Kaskman, R., Zakharov, S., Shugurov, I. & Ilic, S. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0 (2019).

[42] Hodan, T. *et al.* T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 880–888 (IEEE, 2017).

[43] Xiang, Y., Schmidt, T., Narayanan, V. & Fox, D. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199* (2017).

[44] Denninger, M. *et al.* Blenderproc. *arXiv preprint arXiv:1911.01911* (2019).

[45] Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A. & Gross, M. H. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.* **32**, 73–1 (2013).

[46] Miangoleh, S. M. H., Dille, S., Mai, L., Paris, S. & Aksoy, Y. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9685–9694 (2021).

Data availability

All relevant data that support the findings of this work are available from the corresponding author upon reasonable request.

Author contributions

D.Y. conceived the idea and wrote the manuscript. W.S. and H.Y. were involved in developing the proposed algorithm. D.Y. performed the experiments with help from W.S., S.I.K., B.S., C-K. L., and S.M.. H.Y., J.K., J-Y.H, and G.S. contributed to the theoretical investigations. H-S.L. supervised overall works. All authors participated in discussions and contributed to the manuscript.

Competing interests

The authors declare no competing interests.