From picture to 3D holography: End-to-end learning of real-time 3D photorealistic hologram generation from 2D image input

CHENLIANG CHANG,[1,2,5] BO DAI,[1] DONGCHEN ZHU,[2] JIAMAO LI,[2] JUN XIA,[3] DAWEI ZHANG,[1,*] LIANPING HOU,[4] SONGLIN ZHUANG[1]

[1]School of Optical-Electrical and Computer Engineering, Engineering Research Center of Optical Instrument and System, Ministry of Education and Shanghai Key Laboratory of Modern Optics System, University of Shanghai for Science and Technology, Shanghai, 200093, China
[2]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
[3]Joint International Research Laboratory of Information Display and Visualization, School of Electronic Science and Engineering, Southeast University, Nanjing 210096, China
[4]James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK
[5]changchenliang@usst.edu.cn
**Corresponding author: dwzhang@usst.edu.cn*

Received XX Month XXXX; revised XX Month, XXXX; accepted XX Month XXXX; posted XX Month XXXX (Doc. ID XXXXX); published XX Month XXXX

In this letter, we demonstrate a deep-learning based method capable of synthesizing a photorealistic 3D hologram in real-time directly from the input of a single 2D image. We design a fully automatic pipeline to create large-scale datasets by converting any collection of real-life images into pairs of 2D images and corresponding 3D holograms, and train our convolutional neural network (CNN) end-to-end in a supervised way. Our method is extremely computation-efficient and memory-efficient for 3D hologram generation merely from the knowledge of hand-accessible 2D image content. We experimentally demonstrate speckle-free and photorealistic holographic 3D displays from a variety of scene images, opening up a way of creating real-time 3D holography from daily pictures. © 2022 Optical Society of America

Holographic display enables glass-free and true 3D viewing experiences via the modulation of computer-generated hologram (CGH) [1]. CGHs are digitally generated by simulating the physical diffraction propagation from the 3D wave field data to the hologram, but the overall computational efficiency remains rather challenging and limits its application to real-time display.

One primary challenge is the tremendous computational cost required to perform diffraction simulation for holographic wavefront propagation. Several accelerating algorithms for numerical diffraction were designed to tackle this problem [2], such as look-up tables, wavefront-recording plane, multilayer depth discretization, and using graphics processing unit (GPU) computing. Recently, deep-learning based CGH generating strategies [3-8] showed significantly speeding up of runtime for real-time CGH generation as well as high-quality display results for various 3D scenes. However, these methods still heavily depend on the input of the huge amount of volumetric 3D light wave data in the type of multi-depth images [3-6] or RGB-D [7,8], the pre-acquisition and representation of which further hinders the power-efficient CGH generation from original 3D targets.

Creating 3D data for computer-generated holography is another challenge for addressing CGH computational efficiency. 3D cameras, including time-of-flight (TOF) cameras, light-field cameras, stereo cameras, liquid-lens based cameras are employed to acquire dense and deep 3D information [9-11], but all of them have intrinsic limitations in depicting immediate and accurate 3D scene geometry due to the strict devices calibrations and time-consuming post-processing algorithms. A more attractive strategy is to reversely convert a 3D wavefront from a 2D image. Due to the ill-posed nature of this task, attempts at depth estimation from a single image using traditional algorithms are restricted in situations where the particular cue is present [12]. Although deep convolutional neural networks (CNN) can handle better generalization in a more powerful way [13], collecting high-quality and diverse "image + depth" pairs datasets for supervised network training is still a challenging task.

In this work, we propose a data-driven deep convolutional neural network (CNN) framework to enable fully-automatic and

real-time generation of photorealistic 3D CGH end-to-end from a single 2D image input. We enable this pipeline by creating a large-scale CGH dataset with 30000 pairs of 2D images and corresponding ground truth 3D holograms for the supervised training of our CNN model. Our technique includes the first use of 2D natural images as convenient and efficient datasets for the 3D CGH synthesis network. We experimentally demonstrate the robustness of the proposed 3D CGH generation CNN to various natural pictures via high-quality holographic reproduction results.

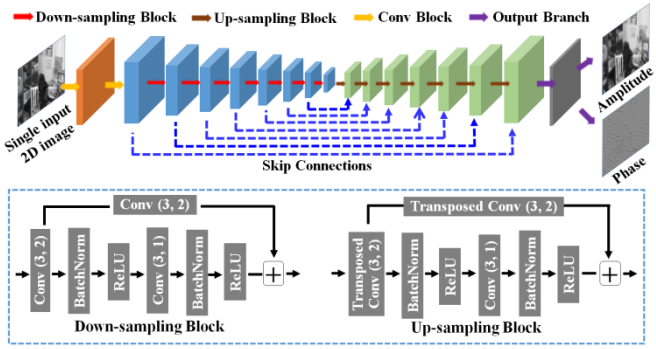

Fig. 1. Schematic of the end-to-end 3D photorealistic hologram generation network architecture. The network receives a single gray-level 2D image as input and rapidly synthesizes a complex hologram that records the corresponding 3D diffractive wavefront of the image content.

Figure 1 shows the structure of the proposed CNN. The network receives a single 2D image in gray-level as the input, and maps the image to a predicted monochromatic 3D complex hologram as output. The network is based on the multiscale ResNets architecture consisting of seven down-sampling residual blocks and seven up-sampling residual blocks. The input image is filtered and passed through down-sampling blocks with resolution decreasing. Each up-sampling block uses skip connections from the down-sampling block's activation output, enabling it to resolve higher-resolution details. At the last of the network, the amplitude and phase of the complex hologram are calculated through an output branch, where we employ 1×1 convolution and sigmoid function for fine-tuning and normalization.

In computer vision community, deep learning has already made success in converting 2D image into 3D content. Many deep learning models can predict accurate depth from 2D image by data-driven training strategy. Based on this fact, our intuition is that given the vast number of training datasets it should be possible to train a deep learning model to predict the 3D diffractive wavefront of 2D image content. To facilitate training CNN for our task, we first create a large training dataset consisting the pairs of 2D images and the corresponding ground truth 3D CGHs. An obvious idea for synthesizing such dataset is to use existing RGB-depth (RGB-D) images or light-field images, from which we can pick up a 2D image and calculate its corresponding ground truth 3D CGH for CNN training. However, large quantities of benchmark RGB-D datasets are not available online and are also expensive to acquire using 3D cameras for novel scene domains. Although it is possible to artificially synthesize large-scale 3D models with specialized graphic software, it lacks the generality to adapt to various scenes with a realistic appearance.

Alternatively, we tackle the problem of creating a large-scale training dataset by introducing a fully automatic 2D image to 3D CGH dataset synthesize approach which only requires a collection of single images. Our approach, which we dubbed "MiDaS-Diffraction-based approach", uses a state-of-the-art (SOTA) off-the-shelf monocular depth estimator, named MiDaS [14], to firstly predict a depth map for each training image. Then we refine this depth prediction along with the image together to further compute a 3D CGH using numerical diffraction algorithms. In doing so, we open up the way to synthesize large numbers of dataset from any 2D image.

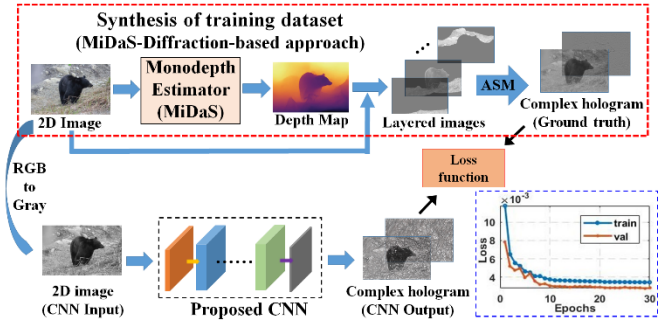

Fig. 2. Workflow for learning 3D holograms from 2D images. We use "MiDaS-Diffraction based approach" to create training dataset consisting of "2D images + ground truth 3D holograms". All the synthesized large-scale data are fed to train the proposed CNN under supervision. ASM: Angular spectrum method

Figure 2 shows one example of synthesizing training dataset using "MiDaS-Diffraction based approach". Starting from any single 2D image of a natural scene, we perform monocular depth estimation using "MiDaS", a network with a ResNeXt-101 backbone, to infer a pixel-level depth map to the 2D image [14]. MiDaS is a state-of-the-art monocular depth estimation network that performs best over all of the other counterparts and is widely used as an off-the-shelf tool for stereo datasets synthesis [15]. MiDaS is trained with a variety of mixed multiple datasets under principled multi-object learning strategy and is quite robust across diverse images. The predicted depth by MiDaS is aligned pixel-by-pixel with the grayscale format of the input 2D image to composite dense 3D volumetric light wave data for monochromatic CGH calculation. According to the point-cloud diffraction model, the complex wavefront of a particular hologram $H(x_m, y_n)$ contributed from all volumetric object points is given by

$$H(x_m, y_n) = \sum_i A(x_i, y_i) \cdot \exp\left[\frac{i2\pi}{\lambda}\sqrt{(x_i - x_m)^2 + (y_i - y_n)^2 + d(x_i, y_i)^2} + i\varphi_0\right], \quad (1)$$

where $A(x_i, y_i)$ is the grayscale amplitude associated with the image pixel (x_i, y_i) at the MiDaS-predicted depth $d(x_i, y_i)$. λ is the wavelength. φ_0 is the position-dependent initial phase to achieve a smooth phase profile. We can further accelerate the calculation of Eq. (1) by dividing the image contents into 256 layer-based depth images according to the 8-bit depth map $d(x, y)$ and calculate the ground truth CGH from each layer image using angular spectrum propagation as

$$H(x_m, y_n) = \sum_{d_i} \mathcal{F}^{-1}\left(\mathcal{F}\left(A(x_i, y_i) \cdot e^{i\varphi_0(x_i, y_i)}\right) \cdot \mathcal{H}_{d_i}(f_x, f_y)\right),$$
$$\mathcal{H}_{d_i}(f_x, f_y) = \begin{cases} e^{i\frac{2\pi}{\lambda}\sqrt{1-(\lambda f_x)^2-(\lambda f_x)^2}\,d_i}, & if\ \sqrt{f_x^2 + f_y^2} < \frac{1}{\lambda} \\ 0 & else \end{cases} \quad (2)$$

where f_x and f_y are the spatial frequencies, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the 2D Fourier transform and inverse Fourier transform operator. d_i (i=1, 2, … 256) is the distance between each layer to the CGH.

A large-scale training dataset consisting of 2D images and corresponding ground truth holograms is synthesized using the MiDaS-diffraction-based approach, and feed to the CNN for end-to-end supervised training with MSE (mean square error) loss function. To maximize the learning ability of our CNN to adapt to different test domains, we generate our training dataset and validation dataset by randomly selecting 30000 and 1000 single images from popular natural image datasets of COCO2017 [16], ADE20K [17], Depth in the Wild [18], and DIODE [19]. We resize all images to 1024×512 pixels and successively apply the monocular depth estimator and diffraction algorithms of Eq. (2) to calculate the corresponding ground truth complex holograms. In our calculation, we set and normalize the 8-bit depth range of all the image scenes from z=0.001m (near plane) to z=0.005m (far plane). The calculation assumes a wavelength of 532nm and a pixel pitch of 6.4 μm. Our CNN is trained on a platform of NVIDIA GeForce RTX 2060 GPU using Adam optimizer with a batch size of 2 and 30 epochs. The learning rate starts at 0.0001 and is decreased by half every 10 epochs. Both of training and validation loss values along with the epochs are also plotted in the lower right of Fig. 2.

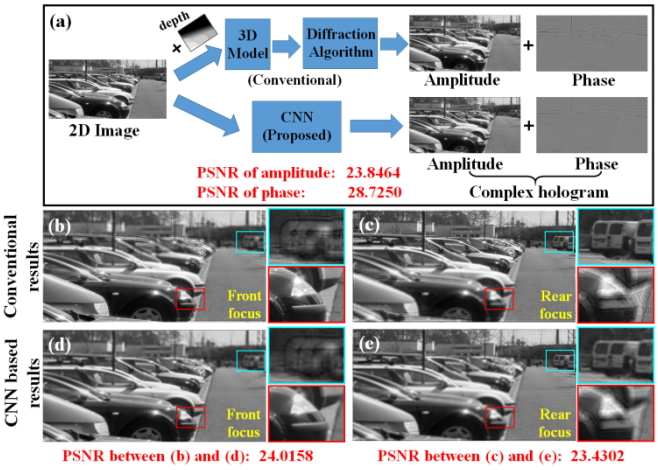

Fig. 3. Simulation results comparing the proposed CNN to the existing algorithm. (a) The generation of ground truth 3D complex hologram from original RGB-D data and the CNN predicted 3D complex hologram from 2D image input. (b)-(c) Simulated reconstructions of front and rear focuses from ground truth hologram. (d)-(e) Simulated reconstructions of front and rear focuses from CNN predicted hologram.

We first quantitatively evaluate our CNN in simulation. Figure 3(a) visualizes a real-world captured RGB-D data where we extract its gray-level 2D image as the CNN input to predict complex hologram. A ground truth hologram is calculated by using Eq. (2) from the original RGB-D volumetric input. Figures 3(b)-3(e) compares the depth-of-field images refocused from the ground truth hologram (Figs. 3(b)-3(c)) and CNN predicted hologram (Fig. 3(d)-3(e)). The results exhibit perceptually similar reconstructions at the front and rear planes. The total runtime of the CNN for inferring hologram from the input image is 0.0175 s on GPU and 0.7661 s on CPU (Intel Core i9-10900), which promises real-time computation performance. We also compute the peak signal-to-noise ratio (PSNR) between CNN and ground truth based results respectively as stated in Fig. 3. Figure 4 shows the experimental setup for the optically holographic 3D display of the CNN predicted hologram. The complex hologram is encoded into a phase-only CGH using a double-phase method [20]. The experiment employs a reflective phase-only SLM (UPOLabs HDSLM64R, 1920×1080, pixel pitch 6.4μm) to display phase-only CGH at a wavelength of 532 nm. The hologram is relayed through a 4-f system where an aperture stop is placed at the Fourier plane to block higher-order diffractions. The reconstructed 3D images are recorded by a digital camera (Nikon D3100) at the rear and far focus depths as shown in Figs. 4(b) and 4(c), which reveals highly consistency with the simulation results of Figs. 3(d) and 3(e).

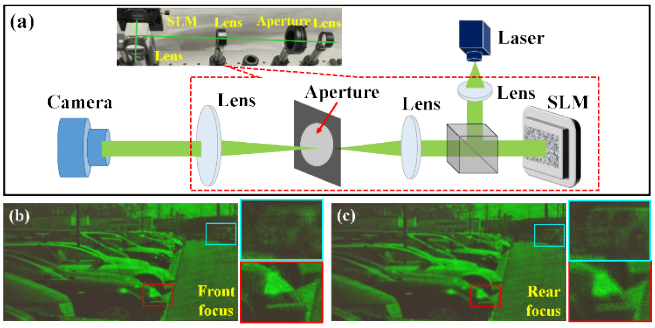

Fig. 4. (a) Experimental setup for holographic 3D display. (b) Optical recordings of the 3D image at front focused distance. (c) Optical recordings of the 3D image at rear focused distance. SLM: spatial light modulator

We further verify our CNN model on a variety of datasets to demonstrate the generality of our CGH generation approach. We choose different natural 2D images of indoor and outdoor scenes from NYU Depth Dataset V2 [21], Make3D Dataset [22], and live-action capture of streetscape using the author's smartphone. We also select several images of synthetic scenes from the SceneFlow dataset [23] and the cartoon movie "Big Buck Bunny" [24]. All of these 2D images are input to our trained CNN and generate corresponded 3D CGHs separately. Figure 5 shows the experimentally reconstructed 3D images recorded at the rear and far focus depths, confirming our CNN-based 3D holography can provide correct focus cues for these 3D targets.

The effective training of our CNN model depends on two aspects: first is the performance of employed monocular depth estimator "MiDaS". We test more cases of 2D image contents where it is challenging to predict depth information using MiDaS. The results are shown and discussed in Supplement 1. Second is the smooth phase distribution assigned to the image contents in the diffraction calculation for ground truth hologram synthesis, which helps boost easy learning of diffraction features with small numerical aperture propagation. But the lack of randomness [25, 26] in such smooth phase strategy degrade the comfort of holographic 3D viewing experiences (see Supplement 1 for more details).

In this paper, we have developed a deep convolutional neural network-based method that can directly and rapidly generate an accurate 3D CGH from the input of a 2D image. The CNN model is trained end-to-end from the large-scale hologram dataset which is synthesized with a fully automatic pipeline from natural images. Our approach enables the possibility of converting any 2D image into 3D CGH in real time, opening up the door to achieve high-quality and realistic 3D holography using immediate capture of 2D pictures from portable and consumer cameras. In the future we intend to explore more functions of the CNN for real-time color 3D holography from single RGB images for potential VR/AR applications.

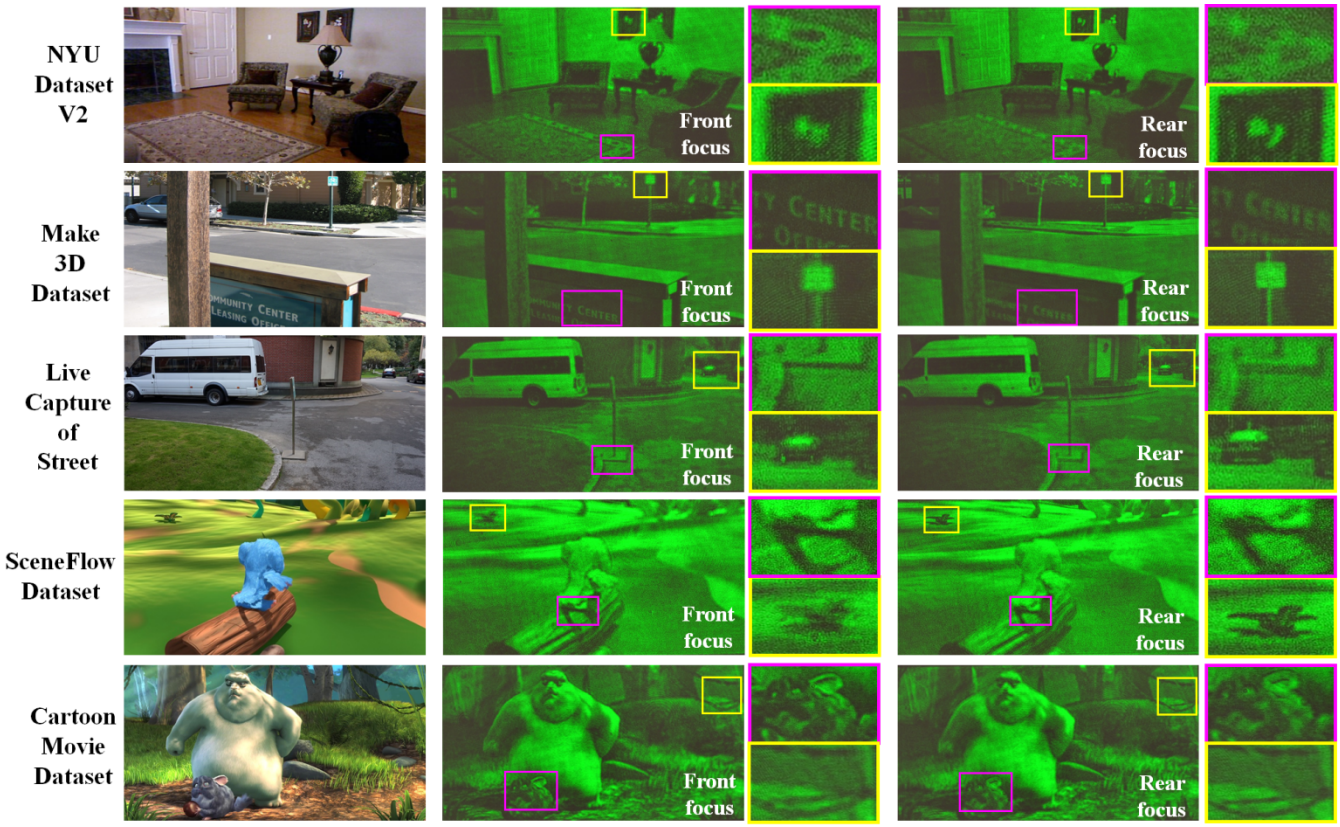

Fig. 5. Holographic 3D display of various scenes from the proposed CNN predicted holograms using 2D image input. The left column shows each input 2D image from various indoor and outdoor image datasets. The middle and right columns are the experimental recordings of 3D reconstructions at front (z=0.001m from the hologram) and real (z=0.005m from the hologram) focal distances.

Funding. Shanghai Municipal Science and Technology Major Project (22ZR1473100); Youth Innovation Promotion Association, Chinese Academy of Sciences (2022232); National Natural Science Foundation of China (62075040); National Key Research and Development Program of China (2021YFF0701100).

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. Y.-L. Li, N.-N. Li, D. Wang, F. Chu, S.-D. Lee, Y.-W. Zheng and Q.-H. Wang, Light: Sci. App. 11, 188 (2022).
2. T. Shimobaba, T. Kakue, and T. Ito, IEEE Trans. Ind. Inf. 12, 1611-1622 (2016).
3. M. H. Eybposh, N. W. Caira, M. Atisa, P. Chakravarthula, and N. C. Pégard, Opt. Express, 28(18), 26636 – 26650 (2020).
4. R. Horisaki, Y. Nishizaki, K. Kitaguchi, M. Saito, and J. Tanida, Appl. Opt. 60(4), A323-A328 (2021).
5. J. Lee, J. Jeong, J. Cho, D. Yoo, B. Lee, and B. Lee, Opt. Express, 28(18), 27137 – 27154 (2020).
6. Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein, ACM Trans. Graph. 39(6), 185 (2020).
7. S. Choi, M. Gopakumar, Y. Peng, J. Kim, and G. Wetzstein, ACM Trans. Graph. 40(6), 240 (2021).
8. L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik, Nature 591, 234-239 (2021).
9. A. G. Marrugo, F. Gao, and S. Zhang, J. Opt. Soc. Am. A 37(9), B60-B77 (2020).
10. D. Wang, C. Liu, C. Shen, Y. Xing, and Q.-H. Wang, PhotoniX, 1(1), 6 (2020).
11. C. Chang, D. Wang, D. Zhu, J. Li, J. Xia, and X. Zhang, Opt. Lett. 47(6), 1482-1485 (2022).
12. Z. He, X. Sui, and L. Cao, Appl. Sci. 11, 9889 (2021).
13. Y. Ming, X. Meng, C. Fan, and H. Yu, Neurocomputing, 438, 14-33 (2021).
14. R, Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, arXiv: 1907.01341 (2019).
15. J. Watson, O. M. Aodha, D. Turmukhambetov, G. J. Brostow, M. Firman, in Proc. of European Conference on Computer Vision, pp. 722-740 (2020).
16. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, in Proc. of European Conference on Computer Vision, pp. 740-755 (2014).
17. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 633-641 (2017).
18. W. Chen, Z. Fu, D. Yang, J. Deng, in Proc. 30th Int. Conf. Neural Inf. Process. Syst., pp. 730-738, (2016).
19. I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, G. Shakhnarovich, arXiv:1908.00463 (2019).
20. Y. Qi, C. Chang, and J. Xia, Opt. Express 24(26), 30368-30378 (2016).
21. N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, In Proceedings of the European conference on Computer Vision (ECCV), pp.746-760, (2012).
22. A. Saxena, M. Sun, and A. Y. Ng, IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 824-840 (2009).
23. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A Dosovitskiy, T. Brox, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4040-4048 (2016).
24. www.bigbuckbunny.org
25. S. Choi, M. Gopakumar, Y. Peng, J. Kim, M. O'toole, and G. Wetzstein, ACM SIGGRAPH Conference Proceedings. (2022).

26. D. Yoo, Y. Jo, S.-W. Nam, C. Chen, and B. Lee, Opt. Lett. 46(19): 4769-4772 (2021).

Full Lists of References

1. Y.-L. Li, N.-N. Li, D. Wang, F. Chu, S.-D. Lee, Y.-W. Zheng and Q.-H. Wang, “Tunable liquid crystal grating based holographic 3D display system with wide viewing angle and large size,” Light: Sci. App. 11, 188 (2022).
2. T. Shimobaba, T. Kakue, and T. Ito, "Review of Fast Algorithms and Hardware Implementations on Computer Holography," IEEE Trans. Ind. Inf. 12, 1611-1622 (2016).
3. M. H. Eybposh, N. W. Caira, M. Atisa, P. Chakravarthula, and N. C. Pégard, “DeepCGH: 3D computer-generated holography using deep learning,” Opt. Express, 28(18), 26636 – 26650 (2020).
4. R. Horisaki, Y. Nishizaki, K. Kitaguchi, M. Saito, and J. Tanida, “Three-dimensional deeply generated holography,” Appl. Opt. 60(4), A323-A328 (2021).
5. J. Lee, J. Jeong, J. Cho, D. Yoo, B. Lee, and B. Lee, “Deep neural network for multi-depth hologram generation and its training strategy,” Opt. Express, 28(18), 27137 – 27154 (2020).
6. Y. Peng, S. Choi, N. Padmanaban, and G. Wetzstein, “Neural holography with camera-in-the-loop training,” ACM Trans. Graph. 39(6), 185 (2020).
7. 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. 40(6), 240 (2021).
8. L. Shi, B. Li, C. Kim, P. Kellnhofer, and W. Matusik, “Towards real-time photorealistic 3D holography with deep neural networks,” Nature 591, 234-239 (2021).
9. A. G. Marrugo, F. Gao, and S. Zhang, "State-of-the-art active optical techniques for three-dimensional surface metrology: a review," J. Opt. Soc. Am. A 37(9), B60-B77 (2020).
10. D. Wang, C. Liu, C. Shen, Y. Xing, and Q.-H. Wang, “Holographic capture and projection system of real object based on tunable zoom lens,” PhotoniX, 1(1), 6 (2020).
11. C. Chang, D. Wang, D. Zhu, J. Li, J. Xia, and X. Zhang, “Deep-learning-based computer-generated hologram from a stereo image pair,” Opt. Lett. 47(6), 1482-1485 (2022).
12. Z. He, X. Sui, and L. Cao, "Holographic 3D Display Using Depth Maps Generated by 2D-to-3D Rendering Approach," Appl. Sci. 11, 9889 (2021).
13. Y. Ming, X. Meng, C. Fan, and H. Yu,“Deep learning for monocular depth estimation: A review,” Neurocomputing, 438, 14-33 (2021).
14. R, Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” arXiv:1907.01341 (2019).
15. J. Watson, O. M. Aodha, D. Turmukhambetov, G. J. Brostow, M. Firman, “Learning stereo from single images,” in Proc. of European Conference on Computer Vision, pp. 722-740 (2020).
16. T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. of European Conference on Computer Vision, pp. 740-755 (2014).
17. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, “Scene parsing through ADE20K dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 633-641 (2017).
18. W. Chen, Z. Fu, D. Yang, J. Deng, “Single-image depth perception in the wild,” in Proc. 30th Int. Conf. Neural Inf. Process. Syst., pp. 730-738, (2016).
19. I. Vasiljevic, N. Kolkin, S. Zhang, R. Luo, H. Wang, F. Z. Dai, A. F. Daniele, M. Mostajabi, S. Basart, M. R. Walter, G. Shakhnarovich, “DIODE: A Dense Indoor and Outdoor DEpth Dataset,” arXiv: 1908.00463 (2019).
20. Y. Qi, C. Chang, and J. Xia, "Speckleless holographic display by complex modulation based on double-phase method," Opt. Express 24(26), 30368-30378 (2016).
21. N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” In Proceedings of the European conference on Computer Vision (ECCV), pp.746-760, (2012).
22. A. Saxena, M. Sun, and A. Y. Ng, "Make3D: learning 3D scene structure from a single still image," IEEE Trans. Pattern Anal. Mach. Intell. 31(5), 824-840 (2009).
23. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A Dosovitskiy, T. Brox, "Large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, " in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4040-4048 (2016).
24. *www.bigbuckbunny.org*
25. 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." ACM SIGGRAPH Conference Proceedings. (2022).
26. D. Yoo, Y. Jo, S.-W. Nam, C. Chen, and B. Lee. "Optimization of computer-generated holograms featuring phase randomness control." Opt. Lett. 46(19): 4769-4772 (2021).