Towards Real-time 3D Computer-Generated Holography with Inverse Neural Network for Near-eye Displays

Wenbin Zhou, Xiangyu Meng, Feifan Qu, Yifan (Evan) Peng
The University of Hong Kong, Hong Kong SAR, China

Abstract
Holography plays a vital role in the advancement of virtual reality (VR) and augmented reality (AR) display technologies. Its ability to create realistic three-dimensional (3D) imagery is crucial for providing immersive experiences to users. However, existing computer-generated holography (CGH) algorithms used in these technologies are either slow or not 3D-compatible. This article explores four inverse neural network architectures to overcome these issues for real time and 3D applications.

Author Keywords
Computer-generated holography; Deep learning; VR/AR; Near-eye displays.

1. Introduction
Near-eye display technology constitutes a pivotal component in various VR/AR wearable devices, providing users with immersive and personalized visual experiences. However, discomfort, dizziness, and unnatural 3D effect are frequently reported by users of existing products, primarily due to vergence-accommodation conflict and bulky optics in traditional solutions. Holography emerges as one of the most promising techniques, not only offering solutions to existing issues but also affording unique advantages such as high peak brightness, power efficiency, and the ability to correct visual aberrations. Unlike conventional display techniques, holographic displays can recover the complete light field information of the target scene, including both amplitude and phase. Such information necessitates pre-encoding in the holograms, either through optical interference or various computer-generated holography (CGH) methods [1]. Subsequently, a spatial light modulator (SLM) is employed to modulate the incident coherent light according to the generated hologram, reconstructing the target 3D scenes through diffraction and interference. Nonetheless, it remains a challenge of the holographic display to show real-time 3D scenes using existing CGH algorithms due to the trade-off between image quality and algorithm run-time, as well as the lack of support to multi-plane display mode.

Existing CGH methods consist of direct and iterative approaches. Direct approaches [2] are computationally efficient but typically utilize coding techniques that combine the amplitude and phase information of the light field in a phase-only hologram. However, it might lead to a degradation in image quality due to the occurrence of higher-order diffractions. While several iterative techniques, such as the Gerchberg–Saxton method [3], Wirtinger Holography [4], and stochastic gradient descent [5], have demonstrated the capability to enhance image quality, they require significant computational time and are not suitable for real-time applications.

Recently, deep learning has been widely investigated to address the limitations inherent in traditional approaches with promising results. Peng et al. [5] and Chakravarthula et al. [6] both obtained image quality comparable to that of conventional iterative methods at 1080P or 4K in real-time when generating 2D holograms. Shi et al. [8] has presented a comparable network architecture for generating efficient 3D holograms in real-time on a smartphone. However, it still relies on coding techniques requiring additional calibration for desired image quality. In this article, we investigate four inverse neural network architectures capable of generating 3D holograms in real-time via self-supervised learning from multi-plane images reconstructed using the angular spectrum method.

2. Pipeline of Inverse Neural Network
We propose a novel framework (see Figure 1) for inverse network training and phase-only hologram synthesis tailored for 3D holographic near-eye displays. The inverse network derives its name because it takes as input the target 3D scenes represented by RGBD images and directly predicts the SLM phase-only holograms. This approach constitutes the inverse process of many forward models, which typically simulate the propagation of the light wave from the SLM phase to the target scenes. Consequently, forward models rely on time-consuming iterative optimizer to generate the SLM phase for a specific target scene.

On the other hand, the phase-only hologram predicted by inverse network can instantaneously drive the SLM, enabling holographic displays to render the desired 3D content in real-time. During the training phase, the phase-only representation is propagated back to the target planes via the forward model to reconstruct the input scenes. Finally, the loss is computed by comparing the input and the reconstructed 3D amplitudes. In the following section, we introduce each component of the inverse network, including RGBD image preprocessing, the self-supervision strategy and the different variations of the inverse network architectures.

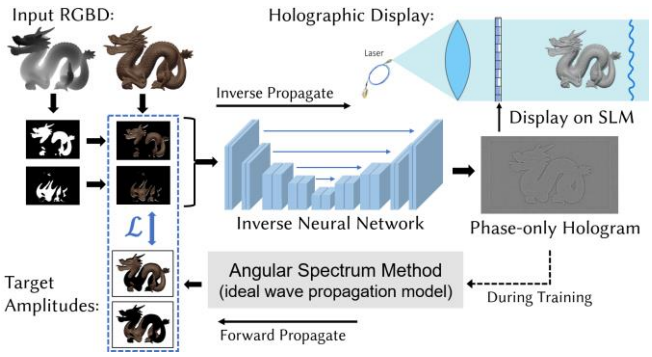

Figure 1 Training and Inference Framework of the Inverse Neural 3D Holography Network: First, the input RGBD images are decomposed into masked images. The inverse neural network then takes masked images as input and outputs a phase-only hologram. During inference, the phase-only hologram drives the SLM inside a holographic display to show target 3D scenes for users. During training, the predicted SLM phase is propagated back via ASM to reconstruct the target multi-plane images. Subsequently, the loss is computed against the masked images and the reconstructed multi-plane images.

ISSN 0097-996X/24/5502-0817-$1.00 © 2024

RGBD image preprocessing: As the representation of target 3D scenes, we adopt RGBD images with three color channels $a_{\text{target}} \in R^{3\times M\times N}$ and one depth channel $D \in R^{M\times N}$. The depth map is then quantized to the nearest target plane, resulting in corresponding masks. For each pixel location, only the pixel on one of the target planes is constrained by the target RGB image, specifically the plane closest to that location according to the depth map. Formally, the mask for each holographic display plane ($j \in [1, J]$) can be expressed as a binary representation $m \in R^{J\times M\times N}$, such that:

$$m^{(j)}(x,y) = \begin{cases} 1 & \text{if } j = \arg\min_k \left|z^{(k)} - D(x,y)\right|, \\ 0 & \text{otherwise .} \end{cases} \quad (1)$$

A pixel in mask location $m^{(j)}(x,y)$ is set to 1 if the depth of that pixel is closest to the j^{th} target plane. It is set to 0 on all other $J-1$ masks, where $z^{(k)}$ is the depth of the k^{th} target plane.

In our framework, a stack of eight masks is computed corresponding to eight target planes which are equally spaced in dioptric space. Specifically, the distance between the target planes and the camera are set to 0.000, 0.084, 0.141, 0.243, 0.317, 0.416, 0.532, and 0.611 $diopters(m^{-1})$. That is due to the ability of human visual system to perceive a maximum of 0.31 diopter inter-plane spacing, which corresponds to the depth of field of human eyes [9].

Therefore, the multi-plane representation can be considered quasi-continuous. After computing the eight masks, we multiply them element-wise with the RGB images to obtain masked images. As we process only one channel of the input RGB image for one inverse network, the eight masks result in an eight-channel grayscale masked image. The masked amplitudes are:

$$a^{(j)}_{\text{masked}}(x,y) = \begin{cases} a_{\text{target}}(x,y) & \text{if } m^{(j)}(x,y) = 1, \\ 0 & \text{if } m^{(j)}(x,y) = 0. \end{cases} \quad (2)$$

Self-supervision: The forward wave propagation model we adopt to reconstruct the target multi-plane images for loss computation is ASM. It is a popular method for computing free-space plane-to-plane wave propagation. As shown in Equation 3, ASM is used to calculate the final wavefront on target planes after the complex wavefront $u(x, y) = a\, e^{i\phi}$ passes through the SLM and propagates in free space by an axial distance z, denoted as:

$$f_{\text{ASM}}(u,z) = \iint \mathcal{F}\left(a(x,y)e^{i\phi(x,y)}\right) e^{i2\pi\left(f_x x + f_Y y + \sqrt{\frac{1}{\lambda^2} - f_X^2 - f_Y^2}\, z\right)} df_X df_Y. \quad (3)$$

We note that ASM is capable of computing both forward and backward wave propagation (by setting z to a negative value). In this project, we adopt ASM to explicitly introduce backward wave propagation in our inverse network and also reconstruct target multiplane amplitudes prior to computing the loss.

Once the target multi-plane amplitudes are reconstructed, the mean squared error loss is computed as $MSE(a_{recon}, a_{target})$ shown in following Equation 4:

$$\text{MSE}\left(a_{\text{recon}}, a_{\text{target}}\right) = \frac{1}{MN}\sum_{j=1}^{8} \left\| m^{(j)}\left(s \cdot a^{(j)}_{\text{recon}} - a_{\text{target}}\right)\right\|_2^2, \quad (4)$$

$s \in R$ is a scaling factor for the reconstructed amplitudes that accounts for possible differences in the range of values between the output of the f_{ASM} and the target amplitude [10].

During training, s can be either fixed or obtained by solving the least-squares problem between the reconstructed and target amplitudes, defined in Equation 5. While in the validation, s is determined by its average value during the last training epoch. We observed that in most of the cases, s converged to a fixed number.

$$s\left(a_{\text{recon}}, a_{\text{target}}\right) = \arg\min_s \left\| m\left(s \cdot a_{\text{recon}} - a_{\text{target}}\right)\right\|_2^2. \quad (5)$$

3. Investigating Inverse Network Architectures

To unlock the ability of 3D multi-plane mode of the holographic near-eye display and achieve real-time performance, we explored several inverse neural network architectures which can be represented as Equation 6. The input and output of the inverse network are 8-channel masked amplitude $a_{masked} \in R^{8\times M\times N}$ and a single-channel SLM phase $\phi_{SLM} \in R^{M\times N}$, respectively.

$$\phi_{SLM} = f_{inverse}(a_{masked}). \quad (6)$$

Single CNN: Firstly, we implemented two distinct convolutional neural network (CNN) architectures. The first one is based on the U-Net model comprising 4 downsampling layers and 4 upsampling layers. The second architecture is solely composed of residual blocks. Each residual block is formed by two convolution layers, two batch normalization layers and two ReLU activation functions. However, a single CNN employed as an inverse neural network only achieves limited image quality under affordable GPU memory cost. It is non-trivial for CNN to learn the cross-domain mapping directly from amplitudes of target 3D scenes to phase-only holograms on an SLM plane [11]. Specifically, the minimal receptive field aggregated across all convolutional layers should at least match the width of the maximum sub-hologram to physically predict the target hologram accurately [7]. This requirement can slow down inference and hinder real-time hologram synthesis.

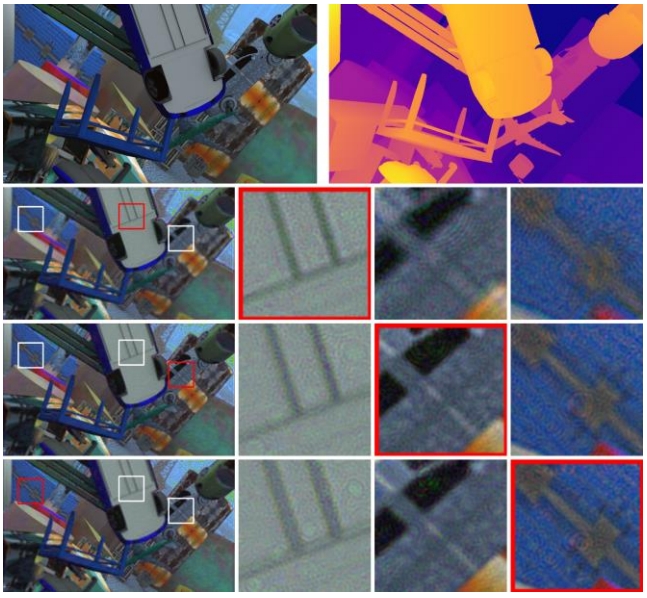

Figure 2 Reconstructed images from a single CNN: The input RGB image and its corresponding depth map (yellow marks a small depth, while purple indicates a large depth) are shown in the first row. Following rows display three reconstructed images on the nearest plane, intermediate plane and farthest plane, respectively. The red boxes denote the in-focus region, whereas the white boxes represent out-of-focus area. It is evident that a single CNN fails to accurately reconstruct proper 3D defocus effects.

CNN-ASM-DPAC: To overcome the problem in the CNN-only model, we propose the Inverse CNN-ASM-DPAC model. The first CNN propagates only to an intermediate plane closer to all the target planes. Then the ASM is adopted to explicitly propagate the intermediate plane wave field back to the SLM plane. Finally, on the SLM plane, the complex-number wave field is encoded into a phase-only hologram by the double phase-amplitude coding (DPAC) method [2]. This approach effectively decomposes the intricate CGH problem into two major parts, cross-domain mapping and long-distance propagation. Long-distance propagation requires a highly deep CNN to solve, hence we adopt the ASM instead. While using ASM to solve cross-domain mapping requires additional iterators, such as stochastic gradient descent, which are notably time-consuming. Consequently, we leverage CNNs to focus on resolving the mapping problem.

Our model is constructed with residual networks and DPAC. The residual network closely resembles the one utilized in CNN-only, but the output is two channels representing the real and imaginary parts of the intermediate plane wave field. DPAC receives the two-channel amplitude and phase, propagated by ASM, and encodes them into a single-channel phase-only hologram.

CNN-ASM-CNN: However, DPAC relies on a coding technique that introduces higher-order sub-diffraction, further diminishing image quality. This arises from the interleaving sampling employed on the phase map, which discards every alternate pixel, resulting in under-sampling for the high-frequency components. One potential solution to mitigate this issue involves incorporating an aperture in the Fourier domain to filter out some of the high-frequency signals. However, this approach necessitates additional optical components and results in a reduction of image luminance.

To address this practical limitation of the DPAC, we propose another CNN to replace it, resulting in a model named Inverse CNN-ASM-CNN. The key distinction lies in the final step, where the phase-only hologram is encoded by CNN instead of DPAC.

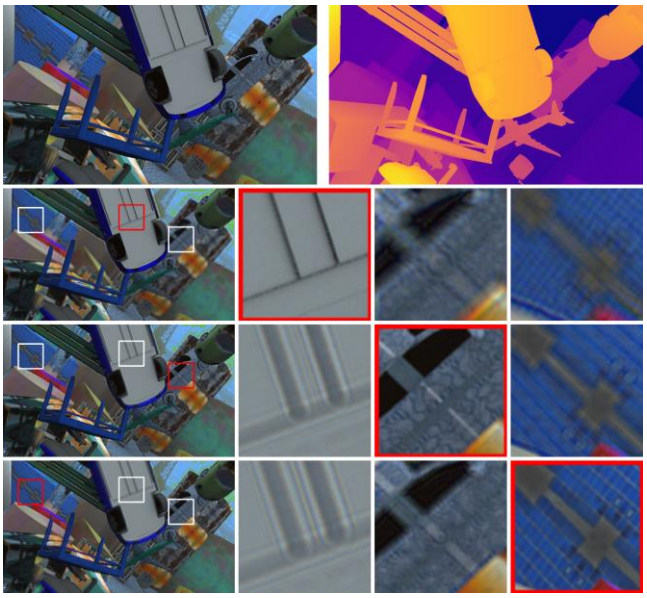

Figure 3 Reconstructed images from CNN-ASM-CNN: The input RGBD image is shown in the first row. Rows 2~4 are the reconstructed images corresponding to the nearest, intermediate and farthest planes. The red boxes denote the in-focus region, while the white boxes represent out-of-focus area, resulting in all sharp images on the diagonal. It shows the objects progressively blur as they move farther away from the focal plane, demonstrating the correct defocus effect.

The inverse CNN-ASM-CNN model consists of two CNNs. The first CNN, referred to as the target-CNN, is a residual network same as the Inverse CNN-ASM-DPAC model. Subsequently, ASM is utilized to propagate the output back to the SLM plane. Finally, our second CNN, named SLM-CNN, adopts the UNet architecture which receives the two-channel propagated amplitude and phase and generates the one-channel phase-only hologram.

Vision Transformer: Recently, vision transformers (ViTs) [12], which have shown leading performance across various computer vision tasks [13], have presented a significant challenge to CNNs. This trend extends to the domain of computer-generated holography. Dong et al. [7] were the first to adopt the ViTs for synthesizing 2D phase-only holograms in real time. Their work demonstrated that 2D inverse neural holography network benefited from ViT's global attention mechanism. We argue that ViT's enhanced ability to model long-range dependencies is essential to simultaneously solve the two major issues encountered in our CNN-only inverse models. Therefore, we propose a ViT-based real-time network for 3D hologram synthesis for the first time.

Specifically, we adopt the U-Former [14] as our inverse network architecture. It consists of 4 down-sampling and up-sampling modules, with each module employs two LeWin Transformer blocks as fundamental components, lowering the computational complexity of vision tasks with high-resolution feature maps. U-Former demonstrates superior capability in capturing local context by integrating a depth-wise convolutional layer between two fully connected layers. Totally, our U-Former receives eight-channel masked images and predicts a one-channel phase-only hologram.

Table 1. Image Quality and Inference Time of Four Inverse Network Architectures

Inverse Network Architecture	PSNR (dB)	Inference Time (s)
Inverse Single CNN	24.8	0.02
Inverse CNN-ASM-DPAC	34.6	0.01
Inverse CNN-ASM-CNN	38.5	0.02
Inverse U-Former	39.1	0.04

4. Results

All inverse neural network architectures were trained and tested with the FlyingThings3D RGBD dataset [15] on a single NVIDIA GeForce RTX 4090 graphics card. The peak signal-to-noise-ratio (PSNR) was used to evaluate the quality of the reconstructed images, and the results are presented in Table 1. Notably, a single CNN failed to generate high-quality phase-only holograms under our experimental conditions. The maximum PSNR of the reconstructed image in the validation set is only around 24.8, and Figure 2 shows that no obvious defocus effect can be observed. In contrast, the inverse CNN-ASM-DPAC architecture achieves a validation PSNR of around 34, with a notable improvement in the visibility of the defocus effect. However, the downside is the additional optical filters required by DPAC in an actual holographic display setup. The best PSNR of the inverse CNN-ASM-CNN architecture on the validation set is 38.54. It demonstrated a pronounced defocus effect on the reconstructed images as shown in Figure 3. Given that U-Former only accepts square images as

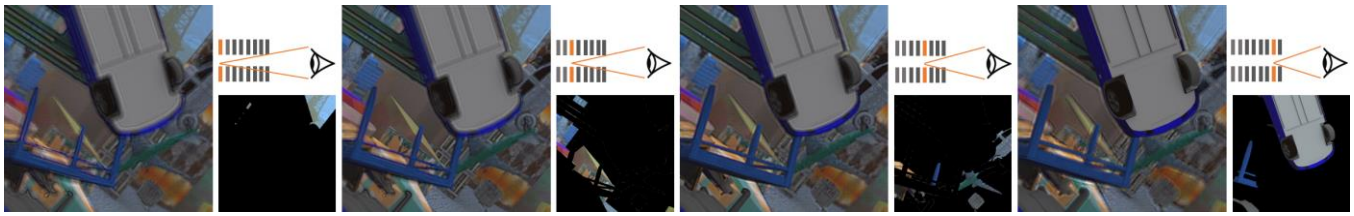

Figure 4 Reconstructed images from U-Former based inverse network: The lower right images show the in-focus region on each target plane. U-Former achieves the best image quality among our 4 inverse networks.

inputs, we initially cropped and resized the RGBD images in the FlyingThings3D dataset to 512×512 to meet the network's requirement. Figure 4 depicts the correct defocus effect on reconstructed images obtained by U-Former, which achieves the highest PSNR among the four inverse architectures.

5. Discussions

This work offers the following insights and contributions:

-- We introduce a self-supervised training framework for the inverse neural 3D holography network that only RGBD images are required to train the network.

-- We conduct an analysis to elucidate why a single CNN fails to directly predict the SLM phase from input RGBD scenes. Specifically, we identify the limited effective receptive field (ERF) is inadequate for capturing the long-range wave propagation.

-- We propose three inverse architectures, namely CNN-ASM-DPAC, CNN-ASM-CNN and U-Former, designed to address the aforementioned challenges. These architectures offer promising improvements on quality and efficiency of 3D hologram synthesis.

We have observed that all three inverse network architectures successfully achieve the correct defocus effect and promising PSNR scores on reconstructed images. This proves that the limited ERF issue can be effectively addressed by either incorporating an ASM operator to explicitly propagate the wave within the model or by leveraging the global attention mechanism of a Vision Transformer. Next, we plan to further improve the quality of reconstructed multiplane images in the out-of-focus region and validate them on optical setups, thereby advancing its applicability in real-world scenarios.

6. Acknowledgements

The authors express their gratitude to Zixuan Wang, Zechen Wang, Zhenyang Li and Rui Wang for their fruitful discussion and support. This work is supported by the Research Grants Council of Hong Kong (ECS 27212822) and the HKU Inno Wing Funding Scheme for student projects/activities by Tam Wing Fan Innovation Fund.

7. References

1. Gopakumar M, Peng Y, Choi S, Kim J, Wetzstein G. 37‐1: Invited Paper: Advances in Neural Holographic Displays for Virtual and Augmented Reality. InSID Symposium Digest of Technical Papers (Vol. 53, No. 1, pp. 454-457), 2022.
2. Maimone A, Georgiou A, Kollin JS. Holographic near-eye displays for virtual and augmented reality. ACM Transactions on Graphics (Tog), 36(4):1-6, 2017.
3. Gerchberg RW. A practical algorithm for the determination of plane from image and diffraction pictures. Optik, 35(2):237-46, 1972.
4. Chakravarthula P, Peng Y, Kollin J, Fuchs H, Heide F. Wirtinger holography for near-eye displays. ACM Transactions on Graphics (TOG), 38(6):1-3, 2019.
5. Peng Y, Choi S, Padmanaban N, Wetzstein G. Neural holography with camera-in-the-loop training. ACM Transactions on Graphics (TOG), 39(6):1-4, 2020.
6. Chakravarthula P, Tseng E, Srivastava T, Fuchs H, Heide F. Learned hardware-in-the-loop phase retrieval for holographic near-eye displays. ACM Transactions on Graphics (TOG), 39(6):1-8, 2020.
7. Dong Z, Xu C, Tang Y, Ling Y, Li Y, Su Y. Vision transformer-based, high-fidelity, computer-generated holography. InAdvances in Display Technologies XIII (Vol. 12443, pp. 47-53), SPIE, 2023.
8. Shi L, Li B, Kim C, Kellnhofer P, Matusik W. Towards real-time photorealistic 3D holography with deep neural networks. Nature, 591(7849):234-9, 2021.
9. Campbell FW. The depth of field of the human eye. Optica Acta: International Journal of Optics, 4(4):157-64, 1957.
10. Choi S, Gopakumar M, Peng Y, Kim J, Wetzstein G. Neural 3D holography: learning accurate wave propagation models for 3D holographic virtual and augmented reality displays. ACM Transactions on Graphics (TOG), 40(6):1-2, 2021
11. Yu T, Zhang S, Chen W, Liu J, Zhang X, Tian Z. Phase dual-resolution networks for a computer-generated hologram. Optics Express, 30(2):2378-89, 2022.
12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
13. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1):87-110, 2022.
14. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: A general u-shaped transformer for image restoration. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 17683-17693), 2022.
15. Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, Brox T. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. InProceedings of the IEEE conference on computer vision and pattern recognition (pp. 4040-4048), 2016.