High-Speed FHD Full-Color Video Computer-Generated Holography

Haomiao Zhang[1,2*], Miao Cao[3,4*], Xuan Yu[5], Hui Luo[6], Yanling Piao[2], Mengjie Qin[2], Zhangyuan Li[1,2], Ping Wang[1,2], Xin Yuan[2†]

[1] Zhejiang University, Hangzhou, Zhejiang, China.
[2] School of Engineering, Westlake University, Hangzhou, Zhejiang, China.
[3] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University.
[4] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.
[5] Guangdong Provincial Key Laboratory of Semiconductor Optoelectronic Materials and Intelligent Photonic Systems, Harbin Institute of Technology, Shenzhen, China.
[6] State Key Laboratory of Optical Field Manipulation Science and Technology, Institute of Optics and Electronics, CAS.
xyuan@westlake.edu.cn

Abstract

Computer-generated holography (CGH) is a promising technology for next-generation displays. However, generating high-speed, high-quality holographic video requires both high frame rate display and efficient computation, but is constrained by two key limitations: (i) Learning-based models often produce over-smoothed phases with narrow angular spectra, causing severe color crosstalk in high frame rate full-color displays such as depth-division multiplexing and thus resulting in a trade-off between frame rate and color fidelity. (ii) Existing frame-by-frame optimization methods typically optimize frames independently, neglecting spatial-temporal correlations between consecutive frames and leading to computationally inefficient solutions. To overcome these challenges, in this paper, we propose a novel high-speed full-color video CGH generation scheme. First, we introduce Spectrum-Guided Depth Division Multiplexing (SGDDM), which optimizes phase distributions via frequency modulation, enabling high-fidelity full-color display at high frame rates. Second, we present HoloMamba, a lightweight asymmetric Mamba-Unet architecture that explicitly models spatial-temporal correlations across video sequences to enhance reconstruction quality and computational efficiency. Extensive simulated and real-world experiments demonstrate that SGDDM achieves high-fidelity full-color display without compromise in frame rate, while HoloMamba generates FHD (1080p) full-color holographic video at over 260 FPS, more than 2.6 times faster than the prior state-of-the-art Divide-Conquer-and-Merge Strategy.

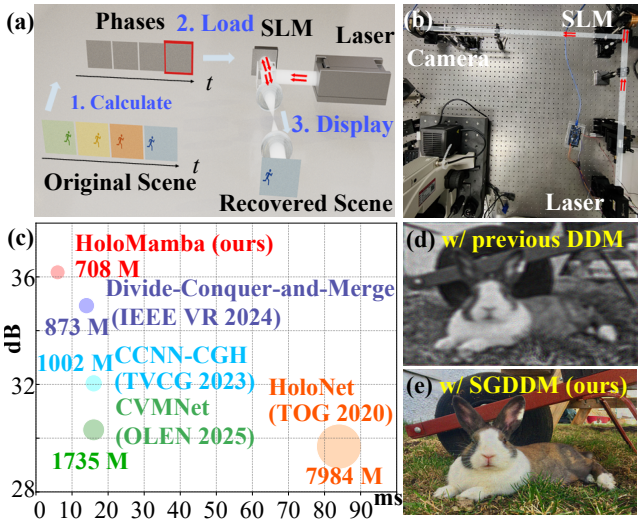

Figure 1: (a) Schematic of CGH model. (b) Real world experimental setup. (c) Comparison of reconstruction quality, testing time and memory requirement with recent deep learning based algorithms. (d,e) Standard DDM fails to perform full-color display with grayscale phase hologram generated by HoloMamba due to the severe color crosstalk. In contrast, our SGDDM preserves high color fidelity.

1 Introduction

Computer-generated holography (CGH) has emerged as an innovative display technology enabling digital synthesis of both real-world and virtual scenes without physical optical constraints (Brown and Lohmann 1966). It holds significant potential in a wide range of applications, including storage (Cheriere et al. 2025), encryption (Fang, Ren, and Gu 2020) and display (Xiong et al. 2021; Gopakumar et al. 2024). As illustrated in Fig. 1(a), a typical CGH pipeline consists of three key steps. First, holograms consisting of phase values are calculated by a reconstruction algorithm from the target intensity information. Next, the holograms are sequentially loaded onto devices such as spatial light modulators (SLMs) to modulate the incident light fields. Finally, the modulated light fields propagate through free space and reconstruct the target information at the display plane, which can be perceived directly by human observers or captured by a camera. Practical CGH displays demand high-speed full-color reconstruction, which requires both

*Equal contribution.
†Corresponding author.

high frame rate display strategies and efficient algorithms.

To achieve *high frame rate full-color display*, depth-division multiplexing (DDM) offers a single-shot solution that operates at the native speed of SLM by superposing phase information of the RGB color channels within one hologram (Makowski, Sypek, and Kolodziejczyk 2008). However, the effectiveness of DDM depends on the optical characteristics of the hologram. Learning-based algorithms tend to generate over-smoothed phase distributions, resulting in limited angular spectral and degraded depth selectivity (Dong et al. 2025). Optically, smooth phase distributions result in an extended depth of field (DOF), which causes reconstructed RGB images to overlap along the axial direction, leading to severe color crosstalk and significant degradation in color fidelity, as shown in Fig.1(d). To eliminate inter-channel crosstalk, time multiplexing (TM) is widely employed, where RGB lasers are sequentially switched, and each is synchronized with its nding hologram displayed on the SLM (Yang et al. 2025; Li et al. 2025). However, TM suffers from a threefold reduction in frame rate, limiting its applicability to dynamic scenes. Therefore, challenge lies in developing a solution that can *simultaneously achieve high temporal resolution and strong color fidelity in full-color holographic* displays.

In parallel, *efficient CGH algorithms* are crucial for high-speed applications. Traditional iterative methods typically rely on techniques such as alternating projection (Gerchberg 1972), stochastic gradient descent (Chen et al. 2019), and non-convex optimization (Candès, Li, and Soltanolkotabi 2015; Candès et al. 2015). These methods are computationally inefficient due to complex optimization procedures. Although existing deep learning approaches have greatly improved the quality and speed of hologram generation, they focus on static hologram generation (Yu et al. 2025). Specifically, CNN-based structures are widely adopted for their ability to model local wavefront efficiently (Zhong et al. 2023b; Li et al. 2025). More recently, Transformer (Dong et al. 2023) and Mamba-based architectures (Yang et al. 2025) have demonstrated strong performance in single-frame CGH tasks. Yet, when extended to *video CGH*, these methods treat each frame independently, which limits efficiency and ignores temporal correlations across frames. Thus a critical challenge remains in designing a *lightweight network* that can efficiently address spatial-temporal modeling for high-speed, high-quality video CGH reconstruction. In other video-related tasks, 3D CNNs suffer from sub-optimal performance due to limited receptive fields (Tran et al. 2015; Chang et al. 2019), and Transformers have high computational costs (Bertasius, Wang, and Torresani 2021; Arnab et al. 2021). In contrast, Mamba employs state space models (SSMs) to capture long-range dependencies with linear complexity, demonstrating high efficiency in long-sequence modeling tasks (Li et al. 2024; Hu et al. 2025), showing promise for real-time video hologram generation.

In a nutshell, we focus on two key design considerations: (i) Suppress color crosstalk in learning-based DDM by reducing phase smoothness and promoting angular spectral diversity across color channels. To this end, we design an effective spectrum-guided modulation strategy dubbed SGDDM in the frequency domain to mitigate color crosstalk, the result is shown in Fig. 1(e). (ii) Design of an efficient spatial-temporal modeling network for high-speed video CGH reconstruction. Motivated by the effectiveness of Mamba in video-related tasks, we propose an efficient network dubbed HoloMamba based on an asymmetric U-Net backbone, integrating both kernel convolution for robust local feature extraction and Mamba modules for efficient long-range spatial-temporal dependency modeling. In this paper, we first adapt and extend this local-global modeling strategy into video CGH, optimizing its capability for comprehensive spatial-temporal representation for high-speed, high-quality reconstruction. Furthermore, we integrate a bidirectional scanning strategy for joint forward and backward temporal modeling. Our main contributions are listed as follows:

- We propose a high-speed, full-color video CGH scheme. Furthermore, we build a *real world holographic display system* based on a phase-only SLM to verify our framework's ability to simultaneously achieve high color fidelity via SGDDM and efficient spatial-temporal video modeling through the HoloMamba network.
- We design an effective spectrum-guided depth division multiplexing (SGDDM), which optimizes phase distribution through frequency modulation and ensures accurate color control, achieving simultaneous full-color holographic display without sacrificing frame rate.
- We propose HoloMamba, an end-to-end lightweight network capable of generating high-speed, high-quality full-color video CGH sequences simultaneously. To our best knowledge, HoloMamba is the first framework to unify efficiency, dynamic spatial-temporal modeling, and high-speed color display in FHD video CGH.

2 Related Work

Full-Color CGH Display

Achieving compact and efficient full-color holographic displays is vital in practical applications. Spatial multiplexing employs three SLMs to improve reconstruction quality but introduces complex optical paths and higher costs (Piao et al. 2019). Time multiplexing synchronizes SLM/laser timing for color display using a single SLM, but triples time consumption (Choi et al. 2022). Frequency multiplexing utilizes one SLM while compromising resolution to simple scenes and requiring additional spatial filtering, thereby increasing optical complexity (Kozacki and Chlipala 2016; Lin, Cao, and Kim 2019). In contrast, depth-division multiplexing (DDM) assigns each color channel (R, G, B) to a distinct focal plane, enabling simultaneous full-color reconstruction on a single SLM (Markley et al. 2023). By avoiding sequential display, DDM preserves temporal resolution and simplifies the optical system, making it well-suited for high frame rate full-color holographic applications. Typically, DDM employs iterative optimization to gradually refine and superimpose the phase information for each RGB channel (Kim and Ee 2023). However, DDM faces a key challenge when combined with deep learning algorithms because networks often produce over-smoothed phase distri-

butions with extended DOF, causing color crosstalk in full-color CGH display.

Efficient CGH Algorithms

Deep learning have greatly improved the quality and efficiency of CGH, marking a significant step toward real-time, high-quality holographic displays (Shui et al. 2022; Liu et al. 2023). Specifically, Peng et al. introduced a novel CGH architecture named HoloNet, which enables real-time 2D holographic displays with approximately 64 ms for FHD hologram generation (Peng et al. 2020). Shi et al. proposed a residual network architecture that efficiently synthesizes photorealistic full-color 3D holograms in about 300 ms (Shi et al. 2021; Shi, Li, and Matusik 2022). Zhong et al. employed lightweight Complex-CNNs (CCNNs) to achieve the same goal, with both approaches reducing GPU memory usage by pruning network parameters, resulting in a generation time of 16 ms for FHD grayscale images (Zhong et al. 2023a). Dong proposed a divide-and-conquer strategy combined with a merging approach to address the challenges of limited memory and computational capacity in CGH generation, with an acceleration of up to 3× and 2× compared with HoloNet and CCNNs respectively (Dong et al. 2024). With long-range modeling capability and relatively high computational efficiency, Mamba-based CVMNet effectively reduces the number of parameters while generating FHD high-quality holograms in just 16 ms (Yang et al. 2025). Yet, these algorithms are inefficient for video hologram generation, making them impractical for real-world applications.

3 Preliminaries of CGH model

As shown in Fig. 1(a), for any given frame of a video sequence in a holographic display system, the nding optical field modulated by the phase-only SLM is $\boldsymbol{u}_{\text{SLM}} = \exp(i\phi(x,y))$, where $\phi(x,y)$ denotes the loaded phase pattern. Then the intensity on the target plane I_{target} is obtained by the angular spectrum method (ASM) as:

$$I_{\text{target}}(x,y) = \left|\mathcal{F}^{-1}\{H(f_x,f_y)\mathcal{F}\{e^{i\phi(x,y)}\}\}\right|^2, \quad (1)$$

where $H(f_x,f_y) = \exp[ikz\sqrt{1-(\lambda f_x)^2-(\lambda f_y)^2}]$ is the transfer function, $k = 2\pi/\lambda$ represents the wave number, and λ is the wavelength; f_x and f_y are the spatial frequencies in the x and y directions respectively. $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ denote the Fourier and inverse Fourier transform respectively. This framework forms the mathematical backbone of CGH reconstruction and supports multiple multiplexing strategies (e.g., TM, DDM) for full-color holography.

Learning-based CGH in Constrained Optimization: CGH is formulated as an ill-posed inverse problem aiming to find a phase distribution $\phi_{\text{opt}}(x,y)$ such that the propagated intensity matches the ground truth image I_{gt}:

$$\phi_{\text{opt}}(x,y) = \arg\min_{\phi} \mathcal{L}\left(I_{\text{target}}, I_{\text{gt}}\right), \quad (2)$$

where $\mathcal{L}$ is the loss function. Notably, this formulation relies solely on the fidelity of intensity, without incorporating explicit constraints on the spectral or other physical properties. With such naive supervision, it often yields over-smoothed phase solutions, especially when the phase is initialized with a uniform (low-frequency) or zero phase (Sui et al. 2024). The relationship between a phase distribution and its angular spectrum can be described by Parseval's theorem:

$$\iint |\nabla\phi(x,y)|^2\, dxdy = \iint \left(f_x^2+f_y^2\right)|\Phi(f_x,f_y)|^2\, df_x df_y, \quad (3)$$

where $\Phi(f_x,f_y)$ is the Fourier spectrum of $\phi(x,y)$. Due to the limited total gradient energy of a smooth phase distribution, the spectral energy must rapidly decay in the high-frequency region. The relationship between diffraction angle θ and spatial frequency can be described by $\sin\theta = \lambda\sqrt{f_x^2+f_y^2}$, indicating that low-frequency components diffract at smaller angles. As a result, the reconstructed intensity remains relatively invariant over an extended axial range with larger DOF, which can be particularly problematic in color holography. When applied to DDM, the wavelength-dependent transfer function $H(f_x,f_y)$ leads to depth replicas. As shown in Fig. 2, a phase optimized for a green-light target at one depth can produce an in-focus replica at another depth under red illumination. In large-DOF systems, these visually indistinguishable replicas remain sharp and overlap with the target across a wide range, causing unavoidable color–depth ambiguity and degrading color fidelity. By displaying each color channel sequentially, TM avoids inter-channel crosstalk at the cost of temporal resolution. More detailed theoretical analysis can be found in Sections 1 and 2 of Supplementary Material (SM).

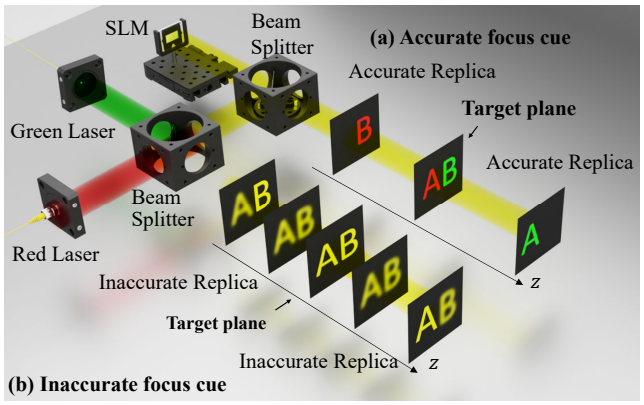

Figure 2: Illustration of depth replica in DDM.

4 Proposed Scheme

In this section, we present a unified scheme composed of SGDDM and HoloMamba, with a particular focus on: (i) guiding the network to expand angular spectrum for high-fidelity color reconstruction without sacrificing temporal resolution, (ii) overcoming inefficiency and redundancy in per-frame CGH optimization by jointly modeling spatial-temporal correlations across video sequences.

Overall Architecture

As illustrated in Fig. 3(a), we treat the intensity of RGB frames as the initial amplitude, while the initial phase is uniformly set to zero, forming a complex-valued input $\boldsymbol{X}_{in} \in$

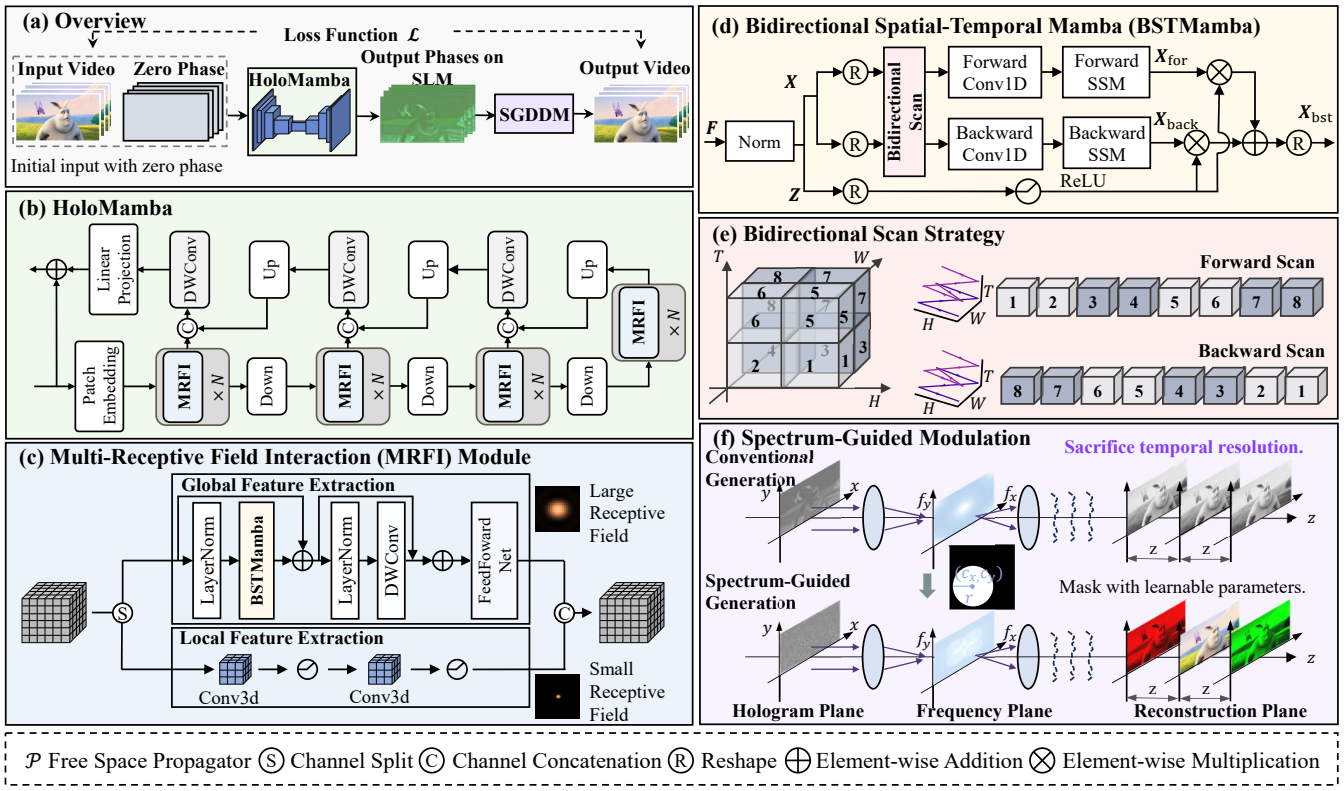

Figure 3: (a) Flowchart of our proposed framework for full-color video CGH reconstruction. (b) The overall network architecture of the proposed HoloMamba. (c) Multi-receptive field interaction (MRFI) module. (d) Bidirectional spatial-temporal mamba (BSTMamba) layer. (e) Bidirectional mamba scanning strategy. (f) Schematic of spectrum-guided modulation.

$\mathbb{C}^{H\times W\times 3\times T}$, where H, W, and T denote the height, width and frame of the video respectively. The phase-only hologram $\boldsymbol{X}_{out} \in \mathbb{C}^{H\times W\times 3\times T}$ is estimated by HoloMamba, a lightweight and efficient reconstruction network that ensures spatial fidelity and temporal coherence across full-color video sequences. The output phase is then passed through our SGDDM to generate the final intensity output.

HoloMamba Structure

We design HoloMamba as a three-level asymmetric U-Net for efficient CGH reconstruction, as shown in Fig.3(b). The patch embedding module adopts a cascaded structure with two 3D convolutions (kernel sizes 3×3×3 and 1×1×1), each followed by LeakyReLU(Maas et al. 2013). A spatial stride of 2 reduces resolution by half while maintaining temporal continuity, enabling a hierarchical projection into a high-dimensional latent space through spatial-temporal filtering. At the decoder output, a linear projection comprising two 3D convolutions (1×1×1 and 3×3×3) reconstructs the final output. The encoder has three layers with multi-receptive field interaction (MRFI) modules and downsampling. MRFI is designed to effectively and efficiently extract features from different scales. The decoder includes three layers with residual depthwise convolution and upsampling. Skip connections are implemented via channel-wise concatenation between nding encoder-decoder layers.

Multi-Receptive Field Interaction Module: As illustrated in Fig. 3(c), MRFI module adopts a hybrid CNN–Mamba architecture, which partitions the input features into two parallel branches along the channel dimension: (i) global receptive feature extraction (C_1 channels) leverages Mamba's linear-complexity state-space modeling for long-range dependency capture. (ii) Local receptive field CNN (C_2 channels) with kernel-level attention to extract spatially localized features. This dual-branch design integrates Mamba's efficient global modeling with CNN's local feature extraction capabilities, achieving a good trade-off between computational efficiency and reconstruction quality.

The local receptive field CNN branch employs a multi-scale 3D convolutional hierarchy to extract local features, operating in parallel with the global receptive branch. Specifically, it applies 3×3×3 and 1×1×1 kernel to balance detail extraction and computational efficiency.

In contrast, the global receptive branch is responsible for modeling long-range spatial-temporal dependencies. As shown in Fig. 3(c), this module is composed of layer normalization, bidirectional spatial-temporal mamba (BSTMamba), depthwise convolution (DWConv), and feed forward network (FFN). We reshape the video data $\boldsymbol{X}^{(k)} \in \mathbb{C}^{B\times C_1\times T\times H\times W}$ into one-dimensional sequences $\boldsymbol{X}^{(k)} \in \mathbb{C}^{B\times C_1\times (T\times H\times W)}$. BSTMamba layer then captures forward-backward contextual relationships using se-

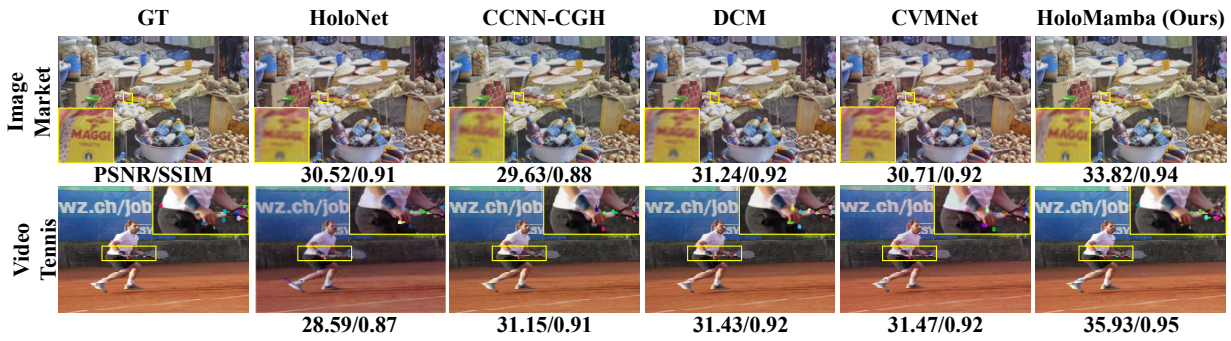

Figure 4: *Numerical Reconstruction Results.* Comparison on full-color FHD image and video datasets. Inter-frame optical flow is visualized for video results to highlight temporal consistency. Zoom in for better view.

lective SSMs, followed by a DWConv layer (kernel size 3×3×3) to preserve fine-grained details. The operations are formulated as:

$$\begin{aligned} \boldsymbol{X}^{(k)} &= \text{BSTM}(\text{LN}(\boldsymbol{X}^{(k-1)})) + \boldsymbol{X}^{(k-1)}, \\ \boldsymbol{X}^{(k)} &= \text{DWConv}(\text{LN}(\boldsymbol{X}^{(k)})) + \boldsymbol{X}^{(k)}, \end{aligned} \tag{4}$$

where LN refers to the layer normalization. The output is then reshaped to $\boldsymbol{X}^{(k)} \in \mathbb{C}^{B\times C_1\times T\times H\times W}$. FFN improves model's non-linearity and its representation ability.

BSTMamba Block: BSTMamba block is designed for efficient spatial-temporal information modeling. As shown in Fig. 3(d), the input feature $\boldsymbol{F}^{(k-1)}$ first passes through two separate linear layers followed by SiLU activations that splits the data into two parallel paths $\boldsymbol{X}^{(k)}$ and $\boldsymbol{Z}^{(k)}$. $\boldsymbol{X}^{(k)}$ is processed by a 1D convolution followed by a ForwardSSM and layer normalization to produce the forward context $\boldsymbol{X}^{(k)}_{\text{For}}$, while a similar pipeline BackwardSSM computes the backward context $\boldsymbol{X}^{(k)}_{\text{Back}}$ using the same $\boldsymbol{X}^{(k)}$. These two directional representations are modulated by $\boldsymbol{Z}^{(k)}$ through element-wise multiplication and summed. The final fused output $\boldsymbol{X}^{(k)}_{\text{bst}}$ is obtained via a linear projection. The process can be formulated as follows, where Lin represents the linear layer and $\odot$ denotes the Hadamard product:

$$\begin{aligned} \boldsymbol{X}^{(k)} &= \text{SiLU}(\text{Lin}(\boldsymbol{F}^{(k-1)})), \\ \boldsymbol{Z}^{(k)} &= \text{SiLU}(\text{Lin}(\boldsymbol{F}^{(k-1)})), \\ \boldsymbol{X}^{(k)}_{\text{For}} &= \text{LN}(\text{ForwardSSM}(\text{Conv1d}(\boldsymbol{X}^{(k)}))), \\ \boldsymbol{X}^{(k)}_{\text{Back}} &= \text{LN}(\text{BackwardSSM}(\text{Conv1d}(\boldsymbol{X}^{(k)}))), \\ \boldsymbol{X}^{(k)}_{\text{bst}} &= \text{Lin}(\boldsymbol{X}^{(k)}_{\text{For}} \odot \boldsymbol{Z}^{(k)} + \boldsymbol{X}^{(k)}_{\text{Back}} \odot \boldsymbol{Z}^{(k)}). \end{aligned} \tag{5}$$

Bidirectional Scan Strategy: As shown in Fig. 3(e), ForwardSSM handles the input 1D sequence along the spatial-temporal axis in the order of Height, Width and Time, while BackwardSSM applies the same operation on the reversed sequence. This design maintains the spatial continuity within each frame while enabling the model to access information from both past and future frames.

Spectrum-Guided Depth-Division Method

In SGDDM, we introduce an explicit spectrum-guided modulation mask $\mathcal{M}_C(f_x, f_y), C \in \{R, G, B\}$ into the Fourier domain as:

$$I_{\text{target}}(x,y) = \left|\mathcal{F}^{-1}\{H(f_x,f_y)\mathcal{M}_C(f_x,f_y)\mathcal{F}\{e^{i\phi(x,y)}\}\}\right|^2, \tag{6}$$

where we recall that the optical field modulated by the phase-only SLM is $\boldsymbol{u}_{\text{SLM}} = \exp(i\phi(x,y))$, with $\phi(x,y)$ being the loaded phase pattern.

As shown in Fig. 3(f), each $\mathcal{M}_C$ is parameterized as a circular binary filter centered at (c_x, c_y) with radius r, where c_x, c_y, and r are learnable parameters. Physically, applying such an off-axis mask in the frequency domain is equivalent to introducing a linear phase ramp in the spatial domain:

$$\phi_{\text{eq}}(x,y) = \phi_{\text{ori}}(x,y) + 2\pi(c_x x + c_y y), \tag{7}$$

where $\phi_{\text{ori}}(x,y)$ denotes the original network output. This modulation shifts the angular spectrum and increases the phase gradient energy:

$$\nabla\phi_{\text{eq}}(x,y) = \nabla\phi_{\text{ori}}(x,y) + 2\pi(c_x, c_y). \tag{8}$$

According to Parseval's theorem in Eq. (3), this spectral shift effectively broadens the angular bandwidth, thereby increasing the numerical aperture (NA) and reducing the depth of field (DOF). For the i-th wavelength, the DOF can be approximated as $\text{DOF}_i \approx \lambda_i/\text{NA}_i^2$. To minimize inter-channel crosstalk, the axial distance between adjacent focal planes should satisfy the following requirement:

$$|z_i - z_j| > \tfrac{1}{2}\left(\tfrac{\lambda_i}{\text{NA}_i^2} + \tfrac{\lambda_j}{\text{NA}_j^2}\right), \quad i \neq j. \tag{9}$$

This condition guides the learning of $\mathcal{M}_C$ to ensure depth-wise separation of RGB channels.

To enable end-to-end training with non-differentiable binary masks, we adopt a continuous surrogate strategy. Specifically, we approximate each binary mask with a soft circular function during training:

$$\widetilde{\mathcal{M}}_C(f_x,f_y) = \sigma\left(\tau\cdot\left[r^2 - (f_x-c_x)^2 - (f_y-c_y)^2\right]\right), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function and τ controls the sharpness of the transition. During the forward pass, the hard binary mask is used to simulate physical masking, while gradients are back-propagated through the soft $\widetilde{\mathcal{M}}_C$ to update the parameters (c_x, c_y, r). We initialize τ at 0.00625 and double it after each training epoch. To ensure numerical stability, we set τ at a maximum value of 1.6 in our experiments. This ensures soft mask to gradually approximate an ideal binary boundary, bridging physical realism and differentiability.

Loss Function

Hereby, we introduce a hybrid loss function designed for holographic reconstruction.

MSE loss: Mean square error (MSE) between the desired video frames $\{\mathbf{y}_t\}_{t=1}^{T} \in \mathbb{R}^{n_x \times n_y}$ and the reconstructed complex amplitude $\hat{\mathbf{x}}_t$ can be written as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{Tn_xn_y} \sum_{t=1}^{T} \|\mathbf{y}_t - |\mathcal{P}(\hat{\mathbf{x}}_t)|\|_2^2, \tag{11}$$

where n_xn_y is the total number of pixels, $\mathcal{P}\{\cdot\}$ denotes the free-space propagation in Eq. (6).

FFL: Focal frequency loss (FFL) emphasizes the recovery of crucial high-frequency components that are often difficult to reconstruct, which is defined as:

$$\mathcal{L}_{\text{FFL}} = \frac{1}{Tn_xn_y} \sum_{t=1}^{T} w \cdot \|\mathcal{F}(\mathbf{y}_t) - \mathcal{F}(|\mathcal{P}(\hat{\mathbf{x}}_t)|)\|_2^2, \tag{12}$$

where $w = |\mathcal{F}(\mathbf{y}_t) - \mathcal{F}(|\mathcal{P}(\hat{\mathbf{x}}_t)|)|^{\alpha}$ is a soft frequency-aware regularization weight, implemented as a dynamic weighting matrix in the Fourier domain. This formulation amplifies the penalty for reconstruction errors in high-frequency regions, and the superscript α controls the sensitivity to frequency-domain discrepancies.

Finally, the overall loss function $\mathcal{L}$ can be written as:

$$\mathcal{L} = \lambda_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + \lambda_{\text{FFL}} \cdot \mathcal{L}_{\text{FFL}}, \tag{13}$$

where λ_{MSE}, λ_{FFL} are constants to balance the two terms. Detailed loss settings are provided in Section 5 of the SM.

5 Experiments

In this section, we evaluate the quality of HoloMamba and SGDDM through simulations and real-world experiments.

Implementation Details

We use Pytorch 2.0 and Python 3.9 trained for 50 epochs with an NVIDIA RTX 8000 GPU. We adopt Adam as the optimizer (Kingma and Ba 2014) with a learning rate of $1e-4$. The `DAVIS2017` dataset (Pont-Tuset et al. 2017) is used to evaluate the trained models. The Peak Signal-to-Noise Ratio (PSNR) and Structured Similarity Index Metric (SSIM) (Wang et al. 2004) are employed to assess the reconstructed results. We perform full-color image experiments on the validation set of `DIV2K` (Lim et al. 2017) and full-color video experiments on the test set of `DAVIS2017`.

Method	$\text{Param}_{(K)}$	$\text{Mem}_{(M)}$	FPS	PSNR/ SSIM	Warp
HoloNet	2868.7	7,984	16	29.69/0.90	0.054
CCNN-CGH	42.2	1,002	61	32.01/0.92	0.032
DCM	112.8	873	99	32.83/0.93	0.048
CVMNet	146.9	1,735	60.1	30.28/0.90	0.031
HoloMamba	**44.7**	**708**	**267↑**	**35.44↑/0.95↑**	**0.022↓**

Table 1: Quantitative comparison of recent CGH algorithms on benchmark datasets with FHD resolution.

Simulation Results

We first compare HoloMamba with HoloNet (Peng et al. 2020), CCNN-CGH (Zhong et al. 2023a), Divide-Conquer-and-Merge Strategy (Dong et al. 2024) (denoted as DCM, using CCNN $\times 4$ configuration), and CVMNet (Yang et al. 2025) using TM for fair comparison. We evaluate both simulated image and video reconstructions. As shown in Fig.1(c) and Tab.1, HoloMamba achieves an average PSNR of 35.44 dB and SSIM of 0.95, outperforming all baselines. In terms of efficiency, our method reduces GPU memory consumption by 18.6% and achieves a 2.6× speedup over DCM. Furthermore, it exhibits lower warping errors compared to other methods in video scenes. Fig. 4 presents selected simulation results on both images and video scenes, where our method demonstrates superior visual quality. The visualization of Lucas-Kanade optical flow shows smoother inter-frame motion trajectories and better structural coherence over time. More simulation results on full-color FHD video sequences are provided in Section 4 of the SM.

We further present full-color holographic results using SGDDM in Fig. 5. Without spectral guidance, HoloMamba produces over-smoothed phases with concentrated low-frequency energy, causing severe color crosstalk. In contrast, SGDDM broadens the spectral distribution, enabling accurate color rendering and high-fidelity reconstruction. Additional results on SGDDM's generalization across baseline networks are provided in Section 4 of the SM.

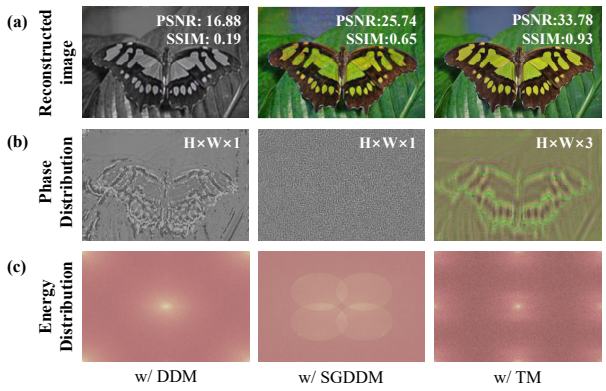

Figure 5: (a) Numerical reconstruction results of full-color images.(b) Phase distribution (c) Light energy distribution of HoloMamba w/ DDM, SGDDM and TM respectively.

Real-World Data Results

Our real data setup is shown in Fig. 1(b), with extra details given in Section 4 of the SM. To fully validate our method's performance, we conduct two key experiments that demonstrate its capabilities across static reconstruction, dynamic video processing, and full-color display applications.

Our first experiment evaluates and compares the performance of HoloMamba against other deep-learning-based real-time algorithms utilizing TM. As shown in Fig. 6, our method produces better reconstruction results, confirming that our network achieves highly effective reconstruction performance while maintaining computational efficiency.

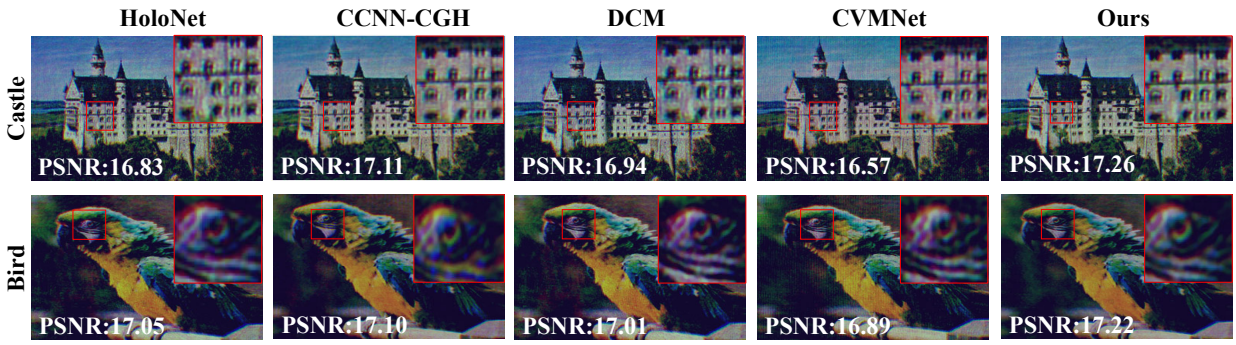

Figure 6: *Real-world data results.* Optical reconstruction results of various algorithms on FHD full-color images based on TM.

Figure 7: *Real-world data results.* Selected reconstruction frames of our HoloMamba on FHD full-color video frames.

Our second experiment demonstrates HoloMamba's capability in large-scale full-color video reconstruction. As shown in Fig. 7, HoloMamba delivers high-quality reconstructions, with the resulting videos exhibiting both high fidelity and smooth temporal transitions across consecutive frames. Additional results are provided in Section 4 of SM.

Ablation Study

We conduct ablation studies on the `DAVIS2017` dataset (256×256 resolution, 8 frames) to validate the effectiveness of HoloMamba. More studies on scanning strategies, architecture and loss design are provided in Section 5 of the SM.

Comparison with CNN and Transformer: We replace HoloMamba with a 3D CNN (Tran et al. 2015) and a vision transformer (ViT) (Dosovitskiy et al. 2020) with global self-attention. From Tab. 2, we can conclude that HoloMamba achieves high performance with computational efficiency. More analysis on the computational cost can be found in Section 3 of SM.

Method	$\text{Mem}_{(\text{M})}$	FPS	PSNR
3D CNN	13,976	17	24.78
ViT	202,348	0.20	29.01
HoloMamba	2,734	267	28.74

Table 2: Ablation study on the HoloMamba model.

MRFI block: We analyze the effect of varying the ratio between global (GRFE) and local (LRFE) receptive field extraction blocks. As shown in Tab. 3, increasing LRFE reduces both parameters and reconstruction quality. Standalone LRFE (0:1) fails to capture long-range spatial-temporal correlations, while standalone GRFE (1:0) improves quality but increases memory usage. Thus, we adopt a 0.8:0.2 ratio to balance efficiency and quality.

$C_1 : C_2$	0:1	0.25:0.75	0.5:0.5	0.8:0.2	1:0
$\text{Mem}_{(\text{M})}$	2,028	2,274	2,499	**2,734**	2,964
PSNR	19.44	20.41	23.28	**28.74**	29.03

Table 3: Ablation study on the MRFI block.

6 Conclusion

We propose a high-speed full-color FHD video CGH scheme to tackle two key challenges: over-smoothed phase that leads to color crosstalk in high frame rate displays, and the lack of spatial-temporal modeling in frame-wise methods. Our proposed SGDDM mitigates color crosstalk by optimizing the angular spectral distribution, enabling one-shot full-color reconstruction without sacrificing frame rate. To further enhance efficiency, we develop HoloMamba, a lightweight asymmetric Mamba-Unet architecture that models spatial-temporal correlations for high-speed phase generation. Extensive simulations and real-world experiments validate the superior performance of our approach, paving a practical way for high frame rate full-color holographic display.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2024YFF0505603, National Natural Science Foundation of China under Grant 62271414, Zhejiang Provincial Science Fund for Distinguished Young Scholar Project under Grant LR23F010001, Zhejiang "Pioneer" and "Leading Goose" R&D Program under Grant 2024SDXHDX0006 and 2024C03182, the Key Project of Westlake Institute for Optoelectronics under Grant 2023GD007, and Ningbo Science and Technology Bureau, "Science and Technology Yongjiang 2035" Key Technology Breakthrough Program under Grant 2024Z126.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, 4.

Brown, B. R.; and Lohmann, A. W. 1966. Complex spatial filtering with binary masks. *Applied optics*, 5(6): 967–969.

Candès, E. J.; Eldar, Y. C.; Strohmer, T.; and Voroninski, V. 2015. Phase retrieval via matrix completion. *SIAM review*, 57(2): 225–251.

Candès, E. J.; Li, X.; and Soltanolkotabi, M. 2015. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4): 1985–2007.

Chang, Y.-L.; Liu, Z. Y.; Lee, K.-Y.; and Hsu, W. 2019. Free-form video inpainting with 3d gated convolution and temporal patchgan. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9066–9075.

Chen, Y.; Chi, Y.; Fan, J.; and Ma, C. 2019. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176: 5–37.

Cheriere, N.; Chu, J.; Brennan, G.; Cameron, P.; Da Costa, P.; Gladrow, J.; Ilunga, G.; Kelly, D.; Lewis, S.; Lim, J.; et al. 2025. Holographic Storage for the Cloud: advances and challenges. *ACM Transactions on Storage*, 21(1): 1–31.

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

Dong, Z.; Jia, J.; Li, Y.; and Ling, Y. 2024. Divide-Conquer-and-Merge: Memory-and Time-Efficient Holographic Displays. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 493–501. IEEE.

Dong, Z.; Ling, Y.; Li, Y.; and Su, Y. 2025. Motion Hologram: Jointly optimized hologram generation and motion planning for photorealistic 3D displays via reinforcement learning. *Science Advances*, 11(5): eads9876.

Dong, Z.; Xu, C.; Tang, Y.; Ling, Y.; Li, Y.; and Su, Y. 2023. Vision transformer-based, high-fidelity, computer-generated holography. In *Advances in Display Technologies XIII*, volume 12443, 47–53. SPIE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, X.; Ren, H.; and Gu, M. 2020. Orbital angular momentum holography for high-security encryption. *Nature Photonics*, 14(2): 102–108.

Gerchberg, R. W. 1972. A practical algorithm for the determination of plane from image and diffraction pictures. *Optik*, 35(2): 237–246.

Gopakumar, M.; Lee, G.-Y.; Choi, S.; Chao, B.; Peng, Y.; Kim, J.; and Wetzstein, G. 2024. Full-colour 3D holographic augmented-reality displays with metasurface waveguides. *Nature*, 1–7.

Hu, X.; Tai, Y.; Zhao, X.; Zhao, C.; Zhang, Z.; Li, J.; Zhong, B.; and Yang, J. 2025. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3581–3589.

Kim, N.; and Ee, H.-S. 2023. Multi-color metasurface hologram based on depth-division multiplexing method. *Journal of the Korean Physical Society*, 82(2): 166–172.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kozacki, T.; and Chlipala, M. 2016. Color holographic display with white light LED source and single phase only SLM. *Optics Express*, 24(3): 2189–2199.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, 237–255. Springer.

Li, Z.-S.; Liu, C.; Li, X.-W.; Zheng, Y.; Huang, Q.; Zheng, Y.-W.; Hou, Y.-H.; Chang, C.-L.; Zhang, D.-W.; Zhuang, S.-L.; et al. 2025. Real-time holographic camera for obtaining real 3D scene hologram. *Light: Science & Applications*, 14(1): 74.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lin, S.-F.; Cao, H.-K.; and Kim, E.-S. 2019. Single SLM full-color holographic three-dimensional video display based on image and frequency-shift multiplexing. *Optics Express*, 27(11): 15926–15942.

Liu, K.; Wu, J.; He, Z.; and Cao, L. 2023. 4K-DMDNet: diffraction model-driven network for 4K computer-generated holography. *Opto-Electronic Advances*, 220135–1.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3. Atlanta, GA.

Makowski, M.; Sypek, M.; and Kolodziejczyk, A. 2008. Colorful reconstructions from a thin multi-plane phase hologram. *Optics express*, 16(15): 11618–11623.

Markley, E.; Matsuda, N.; Schiffers, F.; Cossairt, O.; and Kuo, G. 2023. Simultaneous Color Computer Generated Holography. In *SIGGRAPH Asia 2023 Conference Papers*, 1–11.

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

Piao, Y.-L.; Erdenebat, M.-U.; Kwon, K.-C.; Gil, S.-K.; and Kim, N. 2019. Chromatic-dispersion-corrected full-color holographic display using directional-view image scaling method. *Applied Optics*, 58(5): A120–A127.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.

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

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

Shui, X.; Zheng, H.; Xia, X.; Yang, F.; Wang, W.; and Yu, Y. 2022. Diffraction model-informed neural network for unsupervised layer-based computer-generated holography. *Optics Express*, 30(25): 44814–44826.

Sui, X.; He, Z.; Chu, D.; and Cao, L. 2024. Non-convex optimization for inverse problem solving in computer-generated holography. *Light: Science & Applications*, 13(1): 158.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiong, J.; Yin, K.; Li, K.; and Wu, S.-T. 2021. Holographic optical elements for augmented reality: principles, present status, and future perspectives. *Advanced Photonics Research*, 2(1): 2000049.

Yang, L.; Xu, S.; Yang, C.; Chang, C.; Hou, Q.; and Song, Q. 2025. High-quality computer-generated holography based on Vision Mamba. *Optics and Lasers in Engineering*, 184: 108704.

Yu, X.; Zhang, H.; Zhao, Z.; Fan, X.; Hu, S.; Li, Z.; Chen, W.; Li, D.; Shi, S.; Xiong, W.; et al. 2025. On the use of deep learning for computer-generated holography. *iScience*, 28(5).

Zhong, C.; Sang, X.; Yan, B.; Li, H.; Chen, D.; Qin, X.; Chen, S.; and Ye, X. 2023a. Real-time high-quality computer-generated hologram using complex-valued convolutional neural network. *IEEE Transactions on Visualization and Computer Graphics*.

Zhong, C.; Sang, X.; Yan, B.; Li, H.; Xie, X.; Qin, X.; and Chen, S. 2023b. Real-time 4K computer-generated hologram based on encoding conventional neural network with learned layered phase. *Scientific Reports*, 13(1): 19372.