Learning Coded Apertures for Time-Division Multiplexing Light Field Display

Chun-Hao Chao*, Chang-Le Liu*, and Homer Chen, Fellow, IEEE

Abstract—Conventional stereoscopic displays suffer from vergence-accommodation conflict which cause visual fatigue. Integral imaging-based (II) displays resolves this problem by directly projecting light field sub-views into the eye using microlens arrays. However, II-based light field displays has inherent trade-off between angular and spatial resolutions. In this paper, we propose a novel display concept called coded time-division light field display (C-TDM-LFD), which projects encoded light field sub-views to the viewers' eyes, offering correct cues for vergence-accommodation reflex. By jointly optimizing display inputs and pattern of coded apertures, our pipeline can render high resolution refocused images from sparse light field sub-views with minimal aliasing effects. By simulating light transport and image formation with Fourier optics, we can learn display inputs and coded aperture patterns via deep learning in an end-to-end fashion. To our knowledge, we are among the first to optimize the light field display pipeline with deep learning. We verify our concept with objective image quality metrics (PSNR, SSIM) and optics software simulation, and perform extensive studies on the various customizable design variables in our display pipeline. Experiments results show that our method can generate refocused images with higher quality both quantitatively and qualitatively compared to baseline display designs.

Index Terms—apertures, imaging, three-dimensional displays, optical imaging, optimization.

I. INTRODUCTION

TRADITIONAL stereoscopic displays offer 3-D depth sensation by presenting two images with disparities. Although binocular disparity is one of dominant cue that drives the eye vergence for the human to perceive depth, stereoscopic displays cannot provide correct accommodation cue because the images are projected at a fixed focal distance. Because the human eye is evolved to make its accommodation and vergence consistent, the user usually has eye-strain, headaches, or even depth perception loss after wearing traditional stereoscopic displays, which is referred to as vergence-accommodation conflict (VAC). To address the issue, several 3-D displays have been proposed, such as varifocal displays [1]–[3], multifocal displays [4]–[12], holographic displays [13]–[15], and so on [16]–[19]. However, varifocal displays and multifocal displays both require complicated switchable elements such as eye trackers and screen stacks. On the other hand, holographic displays generally demand high-resolution spatial light modulators.

Among these 3-D displays, integral-imaging based light field display (II-based LFD) [20]–[28] provides 3-D information by projecting different views of a scene to the human eye with a microlens array (MLA). However, one main problem of this type of 3-D display is the spatial-angular resolution trade-off: the spatial resolution of the displayed images is sacrificed to offer higher angular resolution due to the limited display resolution, or vice versa. Low spatial resolution or low angular resolution displays both have degraded image quality due to undersampling.

In this paper, we propose a novel display concept called coded time-division multiplexing light field display (C-TDM-LFD) that leverages coded aperture imaging and time-multiplexing to offer correct cues for vergence-accommodation reflex. By rapidly changing the learned coded aperture patterns in the lens of the display, decoded versions of the display inputs can be projected directly into the viewer's eyes. Through persistence of vision, these projected outputs can then be perceived by the human eye as if they are summed together to generate a refocused image. We show how to model every component in the display pipeline with differentiable functions and optimize the display components with deep learning.

It is worth noting that the term "light field display" in our work refers to display systems where the interval of viewpoints is made smaller than the pupil diameter of eyes. This allows users to view the scene focused at different depths. It should not be confused with displays that allow users to view a scene from different viewpoints (parallax), such as the display proposed by Hirsch et al. [29], which doesn't solve the accommodation problem.

The contributions of our work are highlighted as follows:

- We propose a novel light field display concept called C-TDM-LFD, which is an optimized version of TDM-LFD, to render refocused images from light fields. Experiments show that the quality of refocused images generated by C-TDM-LFD surpasses those generated by II-based LFD.

C.-L. Liu is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: b05901017@ntu.edu.tw).
H. H. Chen is with the Department of Electrical Engineering, Graduate Institute of Communication Engineering, and Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan (e-mail: homer@ntu.edu.tw).

---

*The authors are with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: chunhao.chao@ntu.edu.tw, changleliu@ntu.edu.tw).
TDM-LFD has much higher quality both qualitatively and quantitatively compared to the baseline TDM-LFD, and has natural defocused regions.

- We model every component in the C-TDM-LFD display pipeline with differentiable functions, and optimize the display components via deep learning. Optical software simulation results show that our optimized components can help synthesize high quality refocused images from light fields.
- We use deep learning to learn a mapping from light field sub-view inputs to display inputs prior to the actual display pipeline, which we call the light field encoding Network (LFE). This module transforms light field sub-view inputs to images that are more suitable for light field displays.
- We perform extensive experiments on various design variables (hyper-parameters) in our C-TDM-LFD and show refocused images performance comparisons between different design choices.

II. RELATED WORK

A. Digital Refocusing

Digital refocusing, which generates photographs focused at different depths (distances from the camera) after a single camera shot, has been widely investigated by many researchers in the image processing and computer graphics community. As described in the seminal paper by Levoy et al. on light fields [30], digital refocusing can be simply be approximated as a shift and add operation on light field sub-views. On the other hand, Ng et al [31]. exploited the fact that a photograph is a 2D projection of a 4D light field and used the Fourier Slice Theorem to achieve digital refocusing. However, these refocusing algorithms assume a dense light field is readily available.

To make digital refocusing a more common tool for digital photography, several works have addressed the problem of digital refocusing from a sparse set of views. Bando et al. used blur estimation and deconvolution to generated multiple deconvolved images of an input photograph, and stitched the re-convolved images back together to synthesize a refocused image. The advent of deep learning also gave rise to several learning-based refocusing algorithms. While Dayan et al. [32] used a CNN to directly regress ground truth refocused images, Sakurikar et al. [33] leveraged conditional generative adversarial networks (cGAN) to synthesize refocused images.

Our work is different from previous works on digital refocusing. The refocusing algorithm of previous works either consists of carefully crafted image operations such as blur estimation and deconvolution, or large neural networks. These highly complex operations are hard to be integrated into AR display frameworks. Furthermore, the algorithms in previous works do not consider diffraction, which is often present in most display systems. Simply put, the refocusing step in previous works are simply offline and post-processing computations, which are not suitable for display systems.

Although our goal is indeed to render refocused images for near-eye display, the refocusing operation in our pipeline is purely done by the human eye. By projecting light rays into the human eye, real-time refocusing can be achieved and VAC can be mitigated.

B. Light Field View Synthesis

Densely sampled views from a light field are needed to achieve high quality digital refocusing. However, capturing high spatial and angular resolution light fields is a great challenge, since plenoptic cameras have inherent spatial-angular resolution trade-off and since dense camera arrays are impractical for end-users due to bulkiness and cost.

Many techniques have been developed to address the problem of light field reconstruction from sparsely sampled views, or so-called “angular super-resolution” [34]–[36]. These models aim to increase the angular resolution of the sparsely sampled light fields, e.g. reconstruct a 7×7 light field by interpolating 3×3 sampled views.

In our pipeline, we aim to render refocused images for near-eye displays using a sparse set of light field views. Although conceptually it is possible that we first synthesize a dense light field from sparse light field views and then render a refocused image using our pipeline, it is impractical to do so. In our TDM-based display system, the required frame rate increases linearly with respect to the number of display inputs for each light field to render. Using a full angular resolution light field as display input would lead to physically impossible frame rates. Therefore, we aim to jointly optimize the display inputs and optics pipeline to directly synthesize refocused images for display from sparse light field views.

C. Coded Aperture Imaging

In the past, coded aperture imaging is widely used in the field where X-ray and gamma-ray imaging system [37]–[41]. Because X-ray and gamma-ray are difficult to be refracted or reflected by common optical elements like lens or mirrors, one solution for X-ray and gamma-ray detection is the pinhole camera. However, tiny pinholes account for low light transmittance and hence the low signal-to-noise ratio. To address the problem, the scatter-hole cameras [38] and many coded-aperture cameras [37], [39]–[41], are proposed. These cameras adopt specifically designed apertures to increase the photon collection quantity and the image quality.

Coded aperture imaging is also employed in multiple topics of visible light imaging, like motion deblurring [42], depth estimation [43], high-dynamic range imaging [44], light field capturing [45], [46], and so on [47]–[52]. However, most work focuses on image acquiring. In recent years, the possibility of employing coded-aperture for displays and projectors is gradually explored, like 3-D display [53] and high-quality projector [54].

D. 3-D Display

Most traditional 3-D displays or stereoscopic displays render 3-D information by displaying two different images with disparity to the human eyes, such that the user can sense 3-D objects with eyes convergence. However, in most case, the
shown in also two effective focal lengths behind lengths in front of the lens system such that the image plane is of generality, we place the display panel two effective focal panel, a lens system, and a controllable aperture. Without loss TDMLFD) is outlined. The optical se coded time
field display (TDM
of SDM
disadvantage. The
multiplexing light field display (SDM
rendering different views sample work
 Among these displays, integral-imaging based (II
virtual images are rendered at only one fixed depth. This usually fatigues the user because such rendering mechanism contradicts with human’s accommodation-vergence reflex, which is called VAC. To address the issue, several displays have been proposed, like varifocal display [1]–[3], multifocal display [4]–[12], holographic displays [13]–[15], volumetric displays [55], and light field displays (LFD) [22]–[24], [26], [27], [56]. A more detailed review of these displays can be found in Zabels et al.’s work [57]. Among these displays, integral-imaging based (II-based) light field display (LFD) provides 3-D information by rendering different views sampled from different direction with microlens array (MLA) to the human eye. II-based LFD usually doesn’t need complicate switchable elements and eye trackers that are usually used in varifocal or multifocal displays, or high-resolution spatial light modulators used in holographic display. Several works have proven the possibility of adopting II-based LFD in optical see-through head mounted-display (OST-HMD) [26], [56]. However, one main problem is the low spatial resolution because the perceived spatial resolution equals to the original display resolution divided by the number of lens in MLA. Therefore, in this work, we integrate coded aperture imaging, time-multiplexing, and deep learning techniques into II-based LFD to overcome this problem.

### III. Notation

We consider the 4D light field proposed by Levoy and Hanrahan [30] and define the symbols used in the paper in Table I. Note that s and t are finite in the most cases in practice; therefore, we assume each of them is bounded by N without loss of generality. T

| Table I: NOMENCLATURE |
|------------------------|
| Symbol | Description |
| L(x,y,s,t) | 4-D light field. The pairs (x, y) and (s, t) represent the spatial and angular coordinates, respectively. |
| Ld(x) | Sub-aperture image at s = (x, t). That is, the view captured at the angular coordinates (s, t), where x = (x, y). |
| Lc | Coded light field with \( L_c = \{L_1(x), L_2(x), \ldots, L_N(x)\} \). |
| Pc | Pupil functions represent the coded apertures with \( P_c = \{P_{c1}(x), P_{c2}(x), \ldots, P_{cN}(x)\} \). |
| \( \mathcal{F} \) | Non-unitary Fourier transform. We drop the constant coefficient for simplicity. |
| \( \mathcal{F}^{-1} \) | Inverse Fourier transform. We drop the constant coefficient for simplicity. |

#### IV. Coded Time-Division Light Field Display

In this section, we explain how the traditional space-division multiplexing light field display (SDM-LFD) works and its disadvantage. Then, we describe the time-multiplexed version of SDM-LFD, which is called time-division multiplexing light field display (TDM-LFD). Finally, the setup of the proposed coded time-division multiplexing light field display (C-TDMLFD) is outlined. The optical setup consists of a display panel, a lens system, and a controllable aperture. Without loss of generality, we place the display panel two effective focal lengths in front of the lens system such that the image plane is also two effective focal lengths behind the lens system, as shown in Fig. 1(a).

#### A. Space-Division Multiplexing Light Field Display

Similar to the traditional integral imaging-based light field camera [58], SDM-LFD uses lens array to divide the effective area of the display panel. Each sub-area of the panel displays the corresponding sub-aperture view, and all sub-aperture views overlap on the image plane, as shown in Fig. 1(b). In other words, different directional samplings can be simultaneously perceived at the image plane, and passed onto the human eye by concatenating an eyepiece. Although this design is intuitive, in addition to the manufacturing difficulty of the lens array, one main disadvantage of thus design is the trade-off between the spatial and angular resolutions. For example, for a 3×3 light field display, the actual spatial resolution perceived is only one-ninth of the original one. This makes it almost impossible to mass produce high spatial-resolution light field displays.

#### B. Time-Division Multiplexing Light Field Display

To address the problem of SDM-LFD, one solution is to adopt time-division multiplexing. To construct a time-division multiplexing light field display (TDM-LFD), we put a switchable shutter at the center of the lens system. Note that although only one lens is drawn in Fig. 1 to represent the lens system, the lens system usually consists of multiple lenses (e.g.,
double Gauss lens) so that putting a aperture in is possible. The shutter controls the penetrability of light rays, and is in synchronization with the display panel so that different sub-aperture images of a light field can be displayed by the system in a time-multiplexed fashion, as shown in Fig. 1(c). For example, when the panel displays the sub-aperture image in the top left-hand corner, the corresponding aperture is a shifted small rectangular window in the top left-hand corner.

When the image plane is not two focal lengths behind the lens system, the light rays form a refocused image. The set of refocused images is called the focal stack. The formulation can be derived with Fourier optics [59] as follows.

First, we derive the the formula for our imaging system given on one image. Suppose the complex fields emitted from a monochromatic coherent point source with wave length $\lambda$ on the display panel and received at the image plane are $U_i(x, y)$ and $U_0(x, y)$, respectively. Using Fourier optics, we have

$$
U_i(x, y) = \frac{1}{\lambda^2 z_i z_o} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x_p, y_p) \times \exp \left[ \frac{j}{2} \left( \frac{1}{z_i} + \frac{1}{z_o} - \frac{1}{F} \right) (x^2 + y^2) \right] \times \exp \left[ -j \left( \frac{x_0}{z_o} x_p + \frac{y_0}{z_o} y_p \right) \right] 
\times dx_p dy_p U_0(x_o, y_o) dx_o dy_o
$$

where $h$ is the impulse response of the lens system, $k=\frac{2\pi}{\lambda}$, and $z_0$ and $z_i$ are the object and the image distances, respectively. Note that constant phase and quadratic phase factors have been dropped because they don’t affect the image intensity, on which we focus [59].

Letting $M = -z_i/z_o$ and $x_p = x_0/\lambda z_i$, we have

$$
U_i(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(\lambda z_i x_p, \lambda z_i y_p) \times \exp \left[ \frac{j}{2} \left( \frac{1}{z_o} + \frac{1}{z_i} - \frac{1}{F} \right) (\lambda z_i x_p)^2 + (\lambda z_i y_p)^2 \right] 
\times \exp \left[ -j2\pi((x_0 - M x_o) x_p + (y_0 - M y_o) y_p) \right] 
\times dx_p dy_p U_0(x_o, y_o) dx_o dy_o
$$

That is, $U_i(x, y)$ and $U_0(x, y)$ have the following relationship:

$$
U_i(x, y) = h(x_0, y_0) \ast \frac{1}{|M|} U_0(x_o, y_o)
$$

where $\ast$ is convolution.

Therefore, by assigning different values to $z_i$, we can compute the defocused images of $I_p$. So far, the formula for our optics system is built and we formulate the TDM-LFD in the following paragraph.

Without loss of generality, we consider a light field $L$ taken from a regular grid $G$ of viewpoints on the $(s, t)$-plane with sampling period $\Delta s = (\Delta s, \Delta t)$. For each sub-aperture image $L_{s,t} \in G$, the corresponding aperture can be represented with the shifted rectangular function:

$$
P_s(x) = \text{rect}\left(\frac{x - s \circ \Delta s}{\text{aperture width}}\right)
$$

where $\circ$ is element-wise multiplication. If we let $h_{s,t}$ represent the impulse response with pupil function $P_s$ and image distance $z_i$, then the refocused image at $z_i$ of the TDM-LFD can be represented as follows:

$$h(x_0, y_0)
$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(\lambda z_i x_p, \lambda z_i y_p) 
\times \exp \left[ \frac{j}{2} \left( \frac{1}{z_o} + \frac{1}{z_i} - \frac{1}{F} \right) (\lambda z_i x_p)^2 + (\lambda z_i y_p)^2 \right] 
\times \exp \left[ -j2\pi((x_0 - M x_o) x_p + (y_0 - M y_o) y_p) \right] 
\times dx_p dy_p U_0(x_o, y_o) dx_o dy_o
$$

where $h$ is the impulse response of the lens system, which equals to the Fourier transform of the product of the pupil function $P$ and the aberration function $W$ with given $z_0$.

Because the display is actually incoherent, we need to replace the complex fields $U_i(x, y)$ and $U_0(x, y)$ in Eq. 5 with intensities $I_i(x, y)$ and $I_0(x, y)$:

$$I_i(x, y) = |h(x_0, y_0)| \ast \frac{1}{|M|} I_0(x_o, y_o)
$$

where constant factors are dropped for simplicity. Therefore, by assigning different values to $z_i$, we can compute the defocused images of $I_p$. So far, the formula for our optics system is built and we formulate the TDM-LFD in the following paragraph.
TABLE I

| Symbol | Description |
|--------|-------------|
| $f$    | A non-linear mapping from the light field $L$ to display inputs: this component transforms the light field to the actual images that are input to the light field display. Note that the number of sub-aperture views of the light field and the number of display inputs need not be equal. |
| $\beta$ | A set of coded aperture patterns $\{\beta_1, \beta_2, ..., \beta_k\}$. Display input images are passed through the lens and coded apertures to generate a focal stack. |

TABLE II

| Symbol | Description |
|--------|-------------|
| $n$    | The number of sub-aperture views in the light field |
| $k$    | The number of display input images per light field and the number of coded apertures in the display. It is not necessary that $n = k$. |
| $l$    | The resolution of coded apertures. |
| $q$    | The number of bits used in quantization of the value of coded apertures. |

\[ R_{z_i} = \sum_{s \in G} |h_{s,z_i}|^2 \ast L_s \]  

This setup is a simple way to build a light field display; however, one of the disadvantages is that the apertures allow through little radiation because the aperture width is usually smaller than the sampling period. The other is that aliasing occurs in the defocus regions due to low angular resolution, as shown in Fig. 2.

C. Coded Time-Division Multiplexing Light Field Display

Instead of using traditional grid-like apertures $P$ with light field $L$ as display input, we use specifically engineered apertures $\hat{P}$ and encoded light field $f(L)$ as display input in the proposed C-TDM-LFD, as shown in Fig. 1(c) and Fig. 3. Concretely, we optimize the display input images $f(L)$ and the pupil functions $\hat{P}$ to allow more radiation through the aperture and decrease the aliasing in the bokeh, as shown in Fig. 2. Both $f(L)$ and $\hat{P}$ are optimized simultaneously with deep learning, which we will outline in the following section.

V. OPTIMIZING LIGHT FIELD DISPLAY WITH DEEP LEARNING

In this section, we outline a more implementation-oriented description of our light field display optimization pipeline. To optimize our C-TDM-LFD with deep learning, we model each component in our light field display pipeline with fully differentiable operations.

We optimize two components in our C-TDM-LFD shown in TABLE II, and predefined hyper-parameters shown in TABLE III before the experiments. Note that we call the aperture binary-valued aperture when $q = 2$ and continuous-valued aperture when $q = 8$.

A. Light Field Views to Display Input Mapping

We model the non-linear mapping between the light field views and display input with a convolutional neural network (CNN). We call this network the light field encoding network, denoted by $f$.

Given a sub-aperture views $V_1, V_2, ..., V_n$ of a light field $L$, we concatenate $V_1, V_2, ..., V_n$ into a feature map $M_{in} \in \mathbb{R}^{H \times W \times n}$. We then input $M_{in}$ into our light field encoding network to get $f(M_{in}) = M_{out} \in \mathbb{R}^{H \times W \times k}$. Assuming our C-TDM-LFD takes $k$ encoded light field views as input, we split $M_{out}$ by its last dimension into $k$ images $I_1, I_2, ..., I_n \in \mathbb{R}^{H \times W \times 3}$.

B. Point Spread Function at each Display Plane

In our C-TDM-LFD, the point spread function for a specific image distance $z_i$ is jointly determined by the coded aperture pattern and $z_i$. We denote this point spread function as $h = \mathcal{F}\{\hat{P}_l(\lambda z_i \hat{x}_p, \lambda z_i \hat{y}_p) \cdot W(z_l, \lambda z_l \hat{x}_p, \lambda z_l \hat{y}_p)\} \in \mathbb{R}^{H \times W}$, where $\hat{P}_l \in \mathbb{R}^{l \times l}$ is the $l^{th}$ coded aperture in our C-TDM-LFD. Note that the set of apertures $\hat{P} = \{\hat{P}_1, \hat{P}_2, ..., \hat{P}_k\}$ is fixed for each C-TDM-LFD, and each $\hat{P}_l$ is a variable to optimize in our deep learning framework. We define $l$ as the resolution of the coded aperture.

C. Generating Display Output from Display Input and PSF

According to Section IV, as the display input image $I$ passes through the coded aperture $\hat{P}_l$ and lens, the display generates an output $|h_{\hat{P}_l,I}|^2 \ast I$ at distance $d$ from the lens, where $\ast$ is the 2-D convolution operator.

D. Merging Display Output to Refocused Images with Persistence of Vision

For each light field $L$, we pass sequence of encoded display input images $I_1, I_2, ..., I_k$ through the lens. At each display time step $t$, the display input $I_t$ will be associated with a coded aperture $\hat{P}_t = \hat{P}_{t \mod k}$ (note that the set of apertures $\hat{P} = \{\hat{P}_1, \hat{P}_2, ..., \hat{P}_k\}$ is fixed for each C-TDM-LFD). As the display input switches to the next frame, the coded aperture will also switch to a new pattern. After $k$ time steps, we can see a refocused image at distance $z_i$ from the lens due to persistence of vision, which can be represented as:

\[ R_{z_i} = \sum_{t=1}^{k} |h_{\hat{P}_t,I_t}|^2 \ast I_t \in \mathbb{R}^{H \times W \times 3}. \]

E. Obtaining the Focal Stack

For each light field $L$, we can synthesize a refocused image $R_{z_i}$ at image distance $z_i$ based on the above pipeline. Therefore, we can generate a focal stack $F_{c-tdm} = \{R_{z_{1,1}}, R_{z_{1,2}}, ..., R_{z_{1,m}}\}$ which consists of $m$ images refocused at different depths. To evaluate the quality of $F_{c-tdm}$, we use the focal stack generated from a full angular resolution light field as our ground truth focal stack $F_{gt}$.

F. Losses for Optimization

Because merely optimizing the L1 distance $F_{gt}$ and $F_{c-tdm}$:
means that the reconstruction loss has equal weighting in all pixel locations. In our case, however, we may want to emphasize the reconstruction on certain regions in the refocused image (e.g. in-focus or defocused regions) generated by our C-TDM-LFD. Hence, we propose a weighted reconstruction loss:

$$L_{\text{recon}} = W \odot \| F_{c-\text{tdm}} - F_{gt} \|_1,$$

(8)

where $W \in \mathbb{R}^{M \times H \times W}$ are the weight maps that specify the reconstruction weight for each pixel in each refocused image.

Let $W_j \in \mathbb{R}^{H \times W}$ be the $j^{th}$ weight map in $W$, $1 \leq j \leq m$. To compute $W_j$, we first compute the focus measure response $U_j \in \mathbb{R}^{H \times W}$ from the $j^{th}$ refocused image $R_{z,j}$:

$$U_j = \| \nabla_x R_{z,j} \|_1 + \| \nabla_y R_{z,j} \|_1,$$

(10)

which measures the level of focus of each pixel in $R_{z,j}$, as shown in Fig. 4.

Finally, $W_j$ is acquired by element-wise scaled exponential of the normalized $U_j$:

$$W_j = e^{\beta U_j},$$

$$U_j = \frac{U_j - \min U_j}{\max U_j - \min U_j}$$

(11)

It is obvious that for larger values of $\beta$, more weight is put on in-focus pixels. For $\beta = 0$, $W_j$ is simply a matrix of ones, meaning that every pixel location has the same reconstruction weight.

The full objective of our model is simply:

$$L = L_{\text{recon}}$$

(12)

VI. EXPERIMENTAL SETUP AND RESULTS

In this section, we perform qualitative and quantitative comparisons on two light field datasets. In addition, we verify our concept using the optical design software Zemax by implementing an example design of our proposed C-TDMLFD. Finally, ablation studies are conducted on the various configurations of the display optimization pipeline.

A. Datasets

We verify the effectiveness of our method on two publicly available light field datasets: the INRIA synthetic light field dataset (denoted INRIA-syn) [60] and the INRIA Lytro dataset [61] (denoted INRIA-real).

INRIA-syn dataset consists of 39 light field scenes with 9×9 angular resolution; INRIA-real dataset consists of 63 light field
scenes with $7 \times 7$ angular resolution. We split INRIA-syn and INRIA-real datasets into training/testing data and obtain 31/8 and 51/12 light fields, respectively.

B. Evaluation Metrics

We use two image quality assessment metrics, peak signal-to-noise-ratio (PSNR) and the structural similarity index measure (SSIM), to evaluate the quality of the focal stacks synthesized using our pipeline.

C. Training Details

The light field encoding network $f$ consists of ten stacked residual blocks, as detailed in Fig. 5 and Fig. 6. We clamp the network output to the range $[0, 1]$ with the sigmoid function. For the results shown in the following section, we choose $\beta = 2$ in the weighted reconstruction loss.

All networks are trained end-to-end from scratch for 10000 epochs. We use the Adam optimization algorithm with default parameters $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$, and learning rate 0.001.

D. Quantitative Results

We compare the performance of the focal stack synthesized by C-TDM-LFD and the baseline TDM-LFD where $n = 4$ and $n = 9$, meaning that the display uses four and nine views to render a focal stack. For the four-view setting, the sampled light field views are the four corners views. For the nine-view setting, the sampled light field views form a $3 \times 3$ grid in the original light field, as shown in Fig. 7.

For the four-view setting in our method, we consider CTDM-LFD with $k = 4$. For the nine-view setting, we consider a C-TDM-LFD with $k = 8$.

PSNR and SSIM scores of different settings are summarized in Table IV. We can see that our method generates focal stacks with higher PSNR and SSIM for $n = 4$, $n = 9$ settings on both datasets.

E. Qualitative Results

We also perform visual quality comparisons of the refocused images generated by C-TDM-LFD and the baseline TDM-LFD method. From Fig. 9 and Fig. 14, the refocused images in TDMLFD has severe aliasing artifacts in defocused regions due to under-sampling. On the contrary, our method preserves the sharpness of in-focus regions and has natural blur in defocused regions.

F. Optics Simulation

To verify the practicability of our C-TDM-LFD concept, we use a professional optics design software Zemax to design and simulate an example design. As shown in Fig. 8, we use a unit magnification relay lens in the experiment. The whole system consists of only spherical surfaces and common glass materials (Ohara S-NBM51, S-LAL56, and LAL58), and hence is easy to manufacture. The lens system is symmetric and a binary-valued aperture is put in the center. (Because it is difficult to simulate the continuous-valued aperture in Zemax, we only test binary-
valued aperture in the experiment. The performance of a C-TDM-LFD with binary-valued apertures is slightly lower than a C-TDM-LFD with continuous-valued ones, as discussed in Sec. G.) Note that the proposed C-TDMLFD is not limited to the relay lens. For most imaging or projecting systems, we can similarly put the aperture at the Fourier plane of them.

The simulation result is shown in Fig. 10. We can see that our C-TDM-LFD generates plausible refocused images and that its defocused regions are much more natural compared to refocused images generated with a baseline TDM-LFD.

G. Ablation Studies
In this section, we perform extensive experiments on various design variables (hyper-parameters) in our C-TDM-LFD pipeline and show refocused images performance comparisons between different design choices. All experiments are conducted on the INRIA Synthetic Light Field Dataset.

1) Learned f and \( \hat{P} \):
In the baseline TDM-LFD design in Fig. 1(c), f is simply the identity function. We compare the performance of baseline, only \( \hat{P} \) learned, and both f, \( \hat{P} \) learned. We choose \( n = 4, k = 4, l = 9 \) in this experiment. From Fig. 11 and TABLE IV, we can see that merely optimizing \( \hat{P} \) can improve performance by over 1 dB. Also optimizing the light field encoding mapping f (which is our final C-TDMLFD setting) would further improve performance by nearly 3 dB.

2) Value of coded apertures:
The quantization of the values of coded apertures can be specified by the user depending on the experiment situation. For example, if the coded aperture is implemented by a gray-scale liquid crystal on silicon (LCoS), we can optimize the model with continuous-valued aperture; if the coded aperture is implemented by a binary liquid crystal array, the model is optimized with a continuous-valued aperture.

We experiment on two value configurations for our coded apertures: continuous-valued and binary-valued. For the binary-valued aperture setting, we use the following modified training and inference procedure. During training, \( \hat{P} \) is passed through a sigmoid function with large temperature value \( t \) before computing the point spread functions:

\[
S(x) = \frac{1}{1 + e^{-\alpha}}
\]  

During inference, C is binarized to \{0, 1\} where values larger than 0.5 is replaced with 1 and values less than 0.5 are replaced with 0.

From TABLE VII, we can see that C-TDM-LFD with continuous apertures has better performance. This is very intuitive since the possible patterns that continuous coded apertures can have are almost infinite, which are more flexible for optimization.

3) Symmetry of coded apertures:
Symmetry is sometimes a desirable property in traditional optics design. In our pipeline, we experiment on this concept in the \( k = 4 \) setting by restricting three coded apertures to be mirrored versions of the other one.

TABLE VIII shows that symmetry of coded apertures would lead to degraded performance of refocused images. One possible
reduced number of optimization.

![Fig. 10: Refocused image generated by the proposed C-TDM-LFD.](image)

![Fig. 11: Visual quality comparisons of refocused images generated with C-TDM-LFD with different display components learned. Notice the severe aliasing artifacts in the Baseline and Learned $\hat{P}$ setting is fixed when an additional light field display encoding is learned ($\text{Learned } f, \text{ } \hat{P}$).](image)

![Fig. 12: Visual quality comparisons between refocused images generated with C-TDM-LFD using binary-valued or continuous-valued aperture.](image)

![Fig. 13: Visual comparisons of refocused images generated from models with different $\beta$.](image)

**Table V**

| Method | PSNR↑ / SSIM↑ |
|--------|---------------|
| Synthetic | | |
| Baseline ($n = 4$) | 28.56 / 0.8485 | |
| Ours ($n = 4$, $k = 4$) | 32.29 / 0.9374 | 30.21 / 0.9602 |
| Baseline ($n = 9$) | 34.73 / 0.9688 | 36.32 / 0.9872 |
| Ours ($n = 9$, $k = 8$) | 36.41 / 0.9858 | 39.97 / 0.9891 |

**Table VI**

| Method | PSNR↑ / SSIM↑ |
|--------|---------------|
| Baseline | 28.56 / 0.8485 |
| Learned $\hat{P}$ | 29.85 / 0.8762 |
| Learned $f, \hat{P}$ | 32.29 / 0.9374 |

* Settings: $n = 4$, $k = 4$, $l = 9$, continuous coded aperture.

**Table VII**

| Method | PSNR↑ / SSIM↑ |
|--------|---------------|
| Binary-valued | 31.24 / 0.9249 |
| Continuous-valued | 32.29 / 0.9374 |

* Settings: $n = 4$, $k = 4$, $l = 9$.

**Table VIII**

| Method | PSNR↑ / SSIM↑ |
|--------|---------------|
| Symmetric | 31.06 / 0.9283 |
| Free-valued | 32.29 / 0.9374 |

* Settings: $n = 4$, $k = 4$, $l = 9$.

**Table IX**

| $n$ | PSNR↑ / SSIM↑ |
|-----|---------------|
| 4 | |
| $k = 4$ | 32.29 / 0.9374 |
| $k = 8$ | 33.12 / 0.9304 |
| $k = 16$ | 32.43 / 0.9341 |
| 9 | |
| $k = 4$ | 33.27 / 0.9723 |
| $k = 8$ | 36.41 / 0.9858 |
| $k = 16$ | 34.15 / 0.9795 |

* Settings: $l = 9$.

**Table X**

| $\beta$ | PSNR↑ / SSIM↑ |
|---------|---------------|
| 0 (equal weighting) | 31.25 / 0.9245 |
| 2 | 32.29 / 0.9374 |
| 5 | 32.07 / 0.9376 |

4) Number of light field views used $n$ and encoded light field display inputs $k$:

We experiment on different combinations of $n$, $k$, and evaluate the synthesized refocused image performance. We choose $l = 9$ in this experiment.

From **Table IX**, we can see that larger $n$ has better performance due to increased light field information. However, larger $k$ does not necessarily lead to better performance. In fact, $k$ should also

reason is that capacity of the model is reduced due to the reduced number of optimizable parameters.
not be too large in the practical setting, since the required frame rate of C-TDM-LFD increases linearly with $k$.

5) Weighted reconstruction loss parameter $\beta$:

The $\beta$ parameter in the weighted reconstruction loss controls the relative weighting of reconstruction loss between in-focus and defocus regions in the refocused image. We compare the performance of models with different $\beta$ values.

We found out that global image quality metrics cannot fully reflect the quality of refocused images. From TABLE X and Fig. 13, the global PSNR is highest for $\beta = 2$, yet details in in-focus regions are more preserved for $\beta = 5$. This means that $\beta$ is a hyper-parameter that users can tune to fit their application, depending on how much users want to emphasize on the in-focus regions.

VII. CONCLUSION

In this paper, we propose a novel display concept called coded time-division light field display (C-TDM-LFD), which projects encoded light field sub-aperture views to the viewers’ eyes, offering correct cues for vergence-accommodation reflex. By jointly optimizing display inputs and pattern of coded apertures,
our pipeline can render high resolution refocused images from sparse light field sub-views with minimal aliasing effects. By simulating light transport and image formation with Fourier optics, we can learn display inputs and coded aperture patterns via deep learning in an end-to-end fashion. Theoretical and optical simulation results show that our method can generate refocused images with higher quality both quantitatively and qualitatively compared to baseline display designs.

References

[1] P.-Y. Laffont et al., “Piezo-actuated varifocal head-mounted displays for virtual and augmented reality,” in Advances in Display Technologies IX, San Francisco, United States, Mar. 2019, p. 6. doi: 10.1117/12.2509143.

[2] K. Aksit, W. Lopes, J. Kim, P. Shirley, and D. Luebke, “Near-eye varifocal augmented reality display using see-through screens,” ACM Trans. Graph., vol. 36, no. 6, pp. 1–13, Nov. 2017, doi: 10.1145/3130800.3130892.

[3] A. Wilson and H. Hua, “Design and demonstration of a vari-focal optical see-through head-mounted display using freeform Alvarez lenses,” Opt. Express, vol. 27, no. 11, p. 15627, May 2019, doi: 10.1364/OE.27.015627.

[4] K. Akeley, S. J. Watt, A. R. Girshick, and M. S. Banks, “A stereo display prototype with multiple focal distances,” ACM Trans. Graph., vol. 23, no. 3, pp. 804–813, Aug. 2004, doi: 10.1145/1015706.1015804.

[5] X. Hu and H. Hua, “Design and Assessment of a Depth-Fused Multi-Focal-Plane Display Prototype,” J. Disp. Technol., vol. 10, no. 4, pp. 308–316, Apr. 2014, doi: 10.1109/JDT.2014.2300752.

[6] S. Liu and H. Hua, “A systematic method for designing depth-fused multi-focal plane three-dimensional displays,” Opt. Express, vol. 18, no. 11, p. 11562, May 2010, doi: 10.1364/OE.18.011562.

[7] J. P. Rolland, M. W. Krueger, and A. Goon, “Multifocal planes head-mounted displays,” Appl. Opt., vol. 39, no. 19, p. 3209, Jul. 2000, doi: 10.1364/AO.39.003209.

[8] K. J. MacKenzie, “Vergence and accommodation to multiple-image-plane stereoscopic displays: ‘real world’ responses with practical image-plane separations?”, J. Electron. Imaging, vol. 21, no. 1, p. 011002, Feb. 2012, doi: 10.1117/1.JEI.21.1.011002.

[9] R. Narain, R. A. Albert, A. Bulbul, G. J. Ward, M. S. Banks, and J. F. O’Brien, “Optimal presentation of imagery with focus cues on multi-plane displays,” ACM Trans. Graph., vol. 34, no. 4, pp. 1–12, Jul. 2015, doi: 10.1145/2766909.

[10] J.-H. R. Chang, B. V. K. V. Kumar, and A. C. Sankaranarayanan, “Towards multifocal displays with dense focal stacks,” ACM Trans. Graph., vol. 37, no. 6, pp. 1–13, Jan. 2019, doi: 10.1145/3272127.3275015.

[11] S. Ravikumar, K. Akeley, and M. S. Banks, “Creating effective focus cues in multi-plane 3D displays,” Opt. Express, vol. 19, no. 21, p. 20940, Oct. 2011, doi: 10.1364/OE.19.020940.

[12] K. J. MacKenzie, D. M. Hoffman, and S. J. Watt, “Accommodation to multiple-focal-plane displays: Implications for improving stereoscopic displays and for accommodation control,” J. Vis., vol. 10, no. 8, pp. 22–22, Jul. 2010, doi: 10.1167/10.8.22.

[13] J. F. Heanue, M. C. Bashaw, and L. Hesselink, “Volume Holographic Storage and Retrieval of Digital Data,” Science, vol. 265, no. 5173, pp. 749–752, Aug. 1994, doi: 10.1126/science.265.5173.749.

[14] D. Abookasis and J. Rosen, “Three types of computer-generated hologram synthesized from multiple angular viewpoints of a three-dimensional scene,” Appl. Opt., vol. 45, no. 25, pp. 6533, Sep. 2006, doi: 10.1364/AO.45.006533.

[15] C. Chang, W. Cui, J. Park, and L. Gao, “Computational holographic Maxwellian near-eye display with an expanded eyebbox,” Sci. Rep., vol. 9, no. 1, p. 18749, Dec. 2019, doi: 10.1038/s41598-019-55346-w.

[16] M. Agus, E. Gobbetti, J. A. I. Guitián, F. Marton, and G. Pintore, “GPU Accelerated Direct Volume Rendering on an Interactive Light Field Display,” Comput. Graph. Forum, vol. 27, no. 2, pp. 231–240, Apr. 2008, doi: 10.1111/j.1467-8659.2008.01120.x.

[17] Z. Zhang, Z. Geng, M. Zhang, and H. Dong, “An interactive multimview 3D display system,” San Francisco, California, USA, Mar. 2013, p. 86180P. doi: 10.1117/12.2000360.

[18] Y.-L. Piao, S.-K. Gil, and N. Kim, “Computer-generated holography method based on orthographic projection using depth camera,” in Digital Holography and Three-Dimensional Imaging, Jeju Island, 2017, p. M4B.4. doi: 10.1364/DH.2017.M4B.4.

[19] A. Maimone, A. Georgiou, and J. S. Kollin, “Holographic near-eye displays for virtual and augmented reality,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–16, Jul. 2017, doi: 10.1145/3072959.3073624.

[20] X. Wang and H. Hua, “Time-multiplexed integral imaging based light field displays,” in Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR) II, Mar. 2021, vol. 11765, p. 117650P. doi: 10.1117/12.2576809.

[21] B. Lee, “3D display with floated integral imaging,” SPIE Newsroom, 2007, doi: 10.1117/2.1200702.0547.

[22] H. Arimoto and B. Javidi, “Integral three-dimensional imaging with digital reconstruction,” Opt. Lett., vol. 26, no. 3, pp. 157, Feb. 2001, doi: 10.1364/OL.26.000157.

[23] Y. Taguchi, T. Koike, K. Takahashi, and T. Naemura, “TransCAIP: A Live 3D TV System Using a Camera Array and an Integral Photography Display with Interactive Control of Viewing Parameters,” IEEE Trans. Vis. Comput. Graph., vol. 15, no. 5, pp. 841–852, Sep. 2009, doi: 10.1109/TVCG.2009.30.

[24] H. Hua and B. Javidi, “A 3D integral imaging optical see-through head-mounted display,” Opt. Express, vol. 22, no. 11, p. 13484, Jun. 2014, doi: 10.1364/OE.22.013484.

[25] W. S. Weitao Song, Y. W. Yongtian Wang, D. C. Dewen Cheng, and Y. L. Yue Liu, “Light field head-mounted display with correct focus cue using micro structure array,” Chin. Opt. Lett., vol. 12, no. 6, pp.
The text is a mix of paragraphs introducing authors and their research, followed by lists of publications and a few paragraphs about a person's academic and professional background.

Chang-Le Liu is currently a working toward a M.S degree in communication engineering at National Taiwan University, from 2020. He was a summer intern as a research assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan, and was involved in research in speech enhancement. His current research topic includes image and audio signal processing and light field photography and rendering.

Homer H. Chen (S’83–M’86–SM’01–F’03) received the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana–Champaign. His professional career has spanned industry and academia. Since August 2003, he has been with the College of Electrical Engineering and Computer Science, National Taiwan University, where he is a Distinguished Professor. Prior to that, he held various research and development management and engineering positions at U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island. He was a U.S. delegate for ISO and ITU standards committees and contributed to the development of many new interactive multimedia technologies that are now a part of the MPEG-4 and JPEG-2000 standards. His professional interests lie in the broad area of multimedia signal processing and communications.

Dr. Chen serves on the IEEE Signal Processing Society Awards Board and the Senior Editorial Board of the IEEE Journal on Selected Topics in Signal Processing. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2012 to 2013. He served on the IEEE Signal Processing Society Fourier Award Committee and the Fellow Reference Committee from 2015 to 2017. He was a General Chair of the 2019 IEEE International Conference on Image Processing. He was an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology from 2004 to 2010, the IEEE Transactions on Image Processing from 1992 to 1994, and Pattern Recognition from 1989 to 1999. He served as a Guest Editor for the IEEE Transactions on Circuits and Systems for Video Technology in 1999, the IEEE Transactions on Multimedia in 2011, the IEEE Journal of Selected Topics in Signal Processing in 2014, and Multimedia Tools and Applications (Springer) in 2015.

Chung-Hao Chao is currently working toward a B.S degree in electrical engineering at National Taiwan University, from 2017. He was an undergraduate research assistant at Visual and Learning Lab (VLL), National Taiwan University. He is currently a research assistant at Multimedia Processing and Communications Lab (MPAC) at National Taiwan University and intern at NVIDIA Taiwan. His research interests include image processing, light field processing, and computational photography and displays.

Spectral Imaging,” *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1896–1908, Apr. 2014, doi: 10.1109/TIP.2014.2310125.

[52] A. Mohan, X. Huang, J. Tumblin, and R. Raskar, “Sensing increased image resolution using aperture masks,” in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2008, pp. 1–8. doi: 10.1109/CVPR.2008.4587833.

[53] K. Maruyama, Y. Inagaki, K. Takahashi, T. Fuji, and H. Nagahara, “A 3-D Display Pipeline from Coded-Aperture Camera to Tensor Light-Field Display Through CNN,” in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 1064–1068. doi: 10.1109/ICIP.2019.8803741.

[54] M. Grosse, G. Wetzstein, A. Grundhöfer, and O. Bimber, “Coded aperture projection,” *ACM Trans. Graph.*, vol. 29, no. 3, p. 22:1-22:12, Jul. 2010, doi: 10.1145/1805964.1805966.

[55] G. E. Favalora et al., “100-million-voxel volumetric display,” Orlando, FL, Aug. 2002, pp. 300–312. doi: 10.1117/12.480930.

[56] H. Huang and H. Hua, “High-performance integral-imaging-based light field augmented reality display using freeform optics,” *Opt. Express*, vol. 26, no. 13, p. 17578, Jun. 2018, doi: 10.1364/OE.26.017578.

[57] R. Zabels et al., “AR Displays: Next-Generation Technologies to Solve the Vergence–Accommodation Conflict,” *Appl. Sci.*, vol. 9, no. 15, p. 3147, Aug. 2019, doi: 10.3390/app9153147.

[58] T. Georgiev and C. Intwala, “Light Field Camera Design for Integral View Photography,” Jul. 2008.

[59] J. W. Goodman, *Introduction to Fourier Optics*. Roberts and Company Publishers, 2005.

[60] “Inria synthetic light field datasets.”
http://clim.inria.fr/Datasets/InriaSynLF/index.html (accessed Jul. 03, 2021).

[61] “INRIA Light field dataset.”
http://clim.inria.fr/research/LowRank2/datasets/datasets.html (accessed Jul. 03, 2021).