Research Article Vol. 62, No. 5 / 10 February 2023 / Applied Optics 1337

applied optics

Optical–electronic hybrid Fourier convolutional neural network based on super-pixel complex-valued modulation

Li Fan,[1,†] Xilin Long,[1,†] Jun Dai,[1] Chong Li,[2] Xiaowen Dong,[2] and Jian-Jun He[1,*]

[1]Centre for Integrated Optoelectronics, State Key Laboratory of Modern Optical Instrumentation, College of Optical Science and Engineering, Zhejiang University, Hangzhou 310027, China
[2]Huawei Technologies Co., Ltd., Bantian, Longgang, Shenzhen, Guangdong 518000, China
*Corresponding author: jjhe@zju.edu.cn

Received 17 October 2022; revised 11 January 2023; accepted 12 January 2023; posted 13 January 2023; published 7 February 2023

An optical–electronic hybrid convolutional neural network (CNN) system is proposed and investigated for its parallel processing capability and system design robustness. It is regarded as a practical way to implement real-time optical computing. In this paper, we propose a complex-valued modulation method based on an amplitude-only liquid-crystal-on-silicon spatial light modulator and a fixed four-level diffractive optical element. A comparison of computational results of convolutions between different modulation methods in the Fourier plane shows the feasibility of the proposed complex-valued modulation method. A hybrid CNN model with one convolutional layer of multiple channels is proposed and trained electrically for different classification tasks. Our simulation results show that this model has a classification accuracy of 97.55% for MNIST, 88.81% for Fashion MNIST, and 56.16% for Cifar10, which outperforms models using only amplitude or phase modulation and is comparable to the ideal complex-valued modulation method.

https://doi.org/10.1364/AO.478540

1. INTRODUCTION

The convolutional neural network (CNN) has been widely used in computer vision since AlexNet was demonstrated in 2012 [1,2]. As a type of deep neural network (DNN), CNN consists of feature-extracting convolutional (CONV) layers, feature-merging pooling layers, and fully connected (FC) layers. The convolution is typically conducted using a traditional sliding window spatially moving a kernel matrix across the target matrix. Among all these layers, CONV layers consume the most computing power, especially when it comes to the classification of high-dimensional datasets since they often require a great number of CONV layers to construct a DNN model. Take the example of ResNet, which is now a mainstream scheme in computer vision; its 152-layer prototype has 150 CONV layers taking up most of its total computing power in its inference stage [3]. Despite widespread applications of graphic processing units (GPUs), even more significant computation resources are required for the convolutions of larger images. It remains a significant challenge to reduce the power consumption and latency of DNN models.

A substantial amount of computing power is required in the CONV phase of CNN. Generally, a spatial convolution between an input image of $M \times M$ and a kernel of $N \times N$ requires computation resources proportional to $(M \times M \times N \times N)$. Further, larger images take exponentially longer operation time than smaller ones in both training and inference stages. Consequently, spatial CNN is not viable for large image classification tasks. The Fourier convolutional neural network (FCNN) takes element-wise multiplication in the Fourier domain instead of spatial convolution to accelerate operation speed and maintain excellent performance [4], especially in tasks such as large image classifications.

Owing to the inherent computing parallelism, large bandwidth, and low power consumption of optical and photonic systems, optical computing systems or hardware accelerators have become an area of great interest in recent years [5,6]. There are mainly two categories of optical computing paradigms: all-optical diffraction DNNs (D2NN) and hybrid optical–electronic neural network systems [7,8]. Compared with D2NNs, hybrid systems have reconfigurability and are easier to implement experimentally because of the electronic adaptivity. Using Fourier transformations performed optically by $4f$ systems [9,10], hybrid optical–electronic CNN systems have lower latency, larger bandwidth, and incredible performances. However, most previous hybrid CNN systems implemented convolutions with fixed kernels imposed by a diffractive optical element (DOE) [11,12] or kernels with amplitude-only (AO) Fourier plane modulation [13,14] since commercial spatial light modulators (SLMs) can modulate only either amplitude information (AO) or phase information [phase-only (PO)]

[15]. Complex-valued modulation, which is essential in implementing Fourier convolutions (as shown in Section 2), cannot be performed directly using these SLMs.

There have been many proposed methods to realize complex-valued modulation for computer-generated holography (CGH). The double phase method [16] and phase–amplitude projection method [17] generally employ two SLMs to conduct complex-valued modulation. The digital micromirror device (DMD)-based super-pixel method [18–20] needs an extra spatial low pass filter to compensate for errors brought by its enlarged quantification pitch. Although good performance has been achieved in CGH, these complex-valued modulation methods all add to the system's complexity and cannot be directly applied in Fourier convolutions.

In this paper, we propose a reconfigurable optical–electronic hybrid FCNN model, which involves an AO SLM and a passive four-level DOE to perform super-pixel-based complex-valued modulation in the Fourier plane of the $4f$ system. Multi-channel Fourier convolution is enabled to accelerate CNN's inference phase. Simulation and test results of CNN based on different Fourier-plane modulation methods are presented to demonstrate the advantages of our proposed complex-modulation method. The test accuracy of the FCNN model reaches 97.95%, 88.87%, and 56.16% for MNIST, Fashion MNIST, and Cifar10, respectively, which are all comparable to the ideal complex-valued modulation in the one-layer FCNN model. Theoretical errors caused by the imperfect physical structure of liquid-crystal-on-silicon (LCOS)-based SLMs are also analyzed.

2. HYBRID FCNN ARCHITECTURE

We propose a hybrid optical–electronic FCNN in this section. The architecture of our model is shown in Fig. 1. Fourier convolutions are conducted in optical domain with an optical $4f$ system and a four-level DOE. The derived feature maps are then sent to the computer to complete the classification.

A. Optical System and Modulation Method

1. Optical $4f$ System

In our FCNN architecture, a coherent optical $4f$ system is implemented to operate the Fourier transformation. The optical $4f$ system is a telescope system in which the distance between the image plane and object plane is $4f$, consisting of two convex lenses of the same focal length. The optical field distribution in the back focal plane (Fourier plane) of the first Fourier lens is the Fourier transform of the optical field in the front focal plane (the input SLM) within the Fresnel approximation. Since the spatial-domain convolution is the inverse Fourier transform of the Fourier-domain Hadamard pointwise product, once the input image is loaded onto the object plane while the Fourier transform of the pretrained kernel is loaded onto the Fourier plane, the optically convolved feature map is obtained in the image plane and read out by a CMOS camera, by which instant Fourier convolution is obtained without energy consumption. Reflective nematic-twisted SLMs are employed in our setup; therefore, a polarization beam splitter (PBS) is used with each SLM to realize amplitude modulation. The focal lengths of the lenses in this $4f$ system are both set to 150 mm, and the apertures of these two Fourier lenses are both 12.7 mm to satisfy paraxial Fresnel approximation while maintaining sufficient spatial bandwidth. The spatial bandwidth product is $(D^2/\lambda f)^2$ [21]. For our $4f$ system, the working wavelength is 0.633 μm, so the calculated spatial bandwidth product is approximately 1700×1700, which is adequate for most commercial SLMs.

2. Super-Pixel Complex-Valued Modulation Method

Here we propose a super-pixel scheme for realizing complex-valued modulation using a nematic-twisted-LCOS-based AO SLM with a four-level DOE phase plate attached to its surface. All 2×2 neighboring pixels on the SLM are grouped as one super-pixel to realize complex-valued modulation. Similar super-pixel methods have been proposed in [22,23] using off-axis optical setups. In our scheme, a four-level DOE phase plate is used to induce phase changes in adjacent pixels instead of implementing an off-axis optical setup. The phase plate is implemented on a silica substrate by etching twice in orthogonal directions with depths of $\lambda/4(n-1)$ and $\lambda/8(n-1)$, using the same photomask with a periodic grating pattern [24]. Here n is the refractive index of silica. The physical structure of one super-pixel of this DOE is shown in Fig. 2(a). Each element of the transparent DOE has the same size as the SLM pixel. The 2×2 pixels within a single super-pixel are etched with different etching depths $[0, \lambda/8(n-1), \lambda/4(n-1), 3\lambda/8(n-1)]$ after two etchings, which would, respectively, lead to various

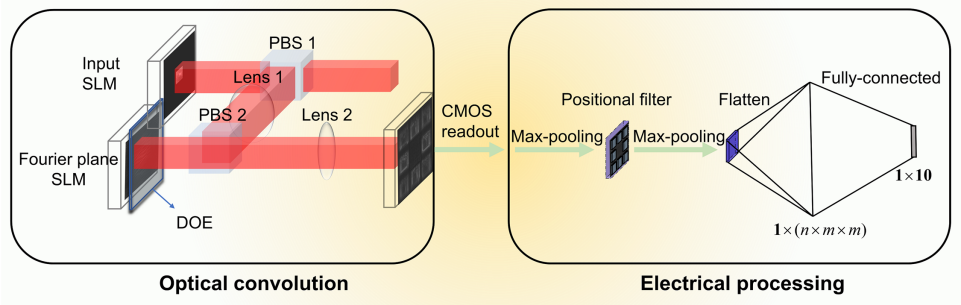

Fig. 1. Optical–electronic hybrid CNN system architecture. The input image is optically convolved. The after-pooling feature maps are element-wise multiplied by a binary positional filter in which the transmission coefficient is unity in areas corresponding to the output feature maps in the output plane and is zero in the padding areas. The filtered data are pooled again and sent into a fully connected layer to get classification results. n, number of convolutional kernels (or output channels); m, original size of the input image.

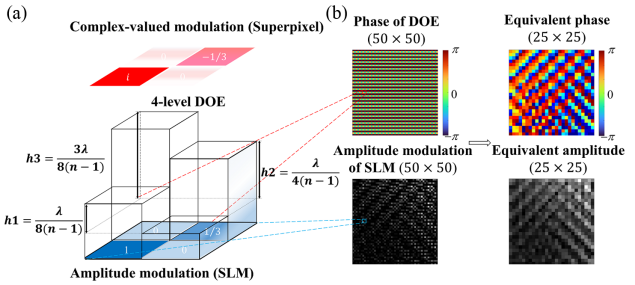

Fig. 2. Illustrations of the proposed super-pixel complex-valued modulation method. (a) Four-level phase modulation of a single super-pixel of the DOE. Four adjacent pixels in the SLM make a super pixel. The DOE has the same pixel size as the SLM; four phase differences of one single super-pixel are induced with four different etching depths in the DOE. n, refractive index of the DOE. Blue pixels: amplitude-only modulation by SLM as the input; red pixels: independent modulation of the real part and imaginary part of a super-pixel after passing twice through the phase plate. (b) 25×25 Fourier-plane complex modulation. Phase distribution of the four-level DOE, modulated amplitude distribution provided by SLM, and the equivalent phase and amplitude distributions are shown.

phase differences $(0, \pi/2, \pi, 3\pi/2)$ since they are passed twice with the reflective SLMs employed. These four phase differences can be regarded as the basic vectors of four directions in the complex plane. By programming the grayscale value of the SLM based on these vectors, any complex-valued vector can be represented. Nematic-twisted LCOS SLM is chosen here to realize 8-bit amplitude modulation. Mathematically, any value in the 2D complex plane can be expressed by the linear combination of two non-collinear basic vectors. For example, the complex value $-1/3+i$ can be modulated with value $+1$ loaded onto the down-left pixel, while value 1/3 is loaded onto the up-right pixel within one super-pixel, as shown in Fig. 2(a). An example of 25×25 complex-valued modulation is shown in Fig. 2(b). We will validate the reliability of performing complex Fourier convolutions with this method in the following sections.

B. Electrical Framework

The electrical backend of the FCNN system is depicted on the right-hand side of Fig. 1. During the inference stage, the resized images are embedded into 600×600 pixel padding images. The trained kernels are tiled onto one 600×600 padding image, and the Fourier transform of the whole image is loaded onto the modulation SLM with super-pixel complex-valued modulation. The optically convolved feature maps are detected by a CMOS array and activated by its inherent saturation function. We simulate this process by implementing tanh as the nonlinear activation function. The convolved results are pooled using max-pooling and extracted by the spatial binary filter, which has unity transmission at the pixels corresponding to the expected positions of the output feature maps and zero transmission at all other pixels. The split results are max-pooled again, and flattened to 1D vectors, and then sent into a FC layer, where it is converted into logit vectors and used to complete the classification tasks.

The training of the hybrid FCNN model includes two electrical training phases. During the pre-training phase, a modified LeNet model containing n CONV kernels is pre-trained to obtain kernels for extracting spatial-invariant features. Since the optical Fourier convolution is a spatially continuous computation, the "stride" of the CONV layer is set to one to simulate this physical process. After that, during the post-training phase, the images of the original dataset are all resized to four times their original size in each dimension and padded to 600×600. The pre-trained kernels are tiled onto a 150×150 padding image and also resized in accordance with the input image to be Fourier-transformed altogether. More details of this preprocessing procedure are discussed in Section 3. With this Fourier-domain pattern loaded into the Fourier plane in different ways (amplitude and phase, AO, PO, and super-pixel methods), the parameters in the FC layers are updated with the stochastic gradient descent (SGD) algorithm, and different optical–electronic FCNN models are trained.

Table 1. Parameters of the Pretrained and FCNN Models[a]

	Pretrained Model	Hybrid FCNN Model
Optical frontend		Fourier conv: size = 600×600 (element wise) Activation function: tanh
Electrical backend	Conv: size = $1 \times n \times h \times w$ stride = 1 Activation function: ReLU Maxpooling: size = 2×2 stride = 2 FC1: $n \times m \times m/4$, 1280 FC2: 1280, 10	Maxpooling1: size = 2×2 stride = 2 Positional filter: size = 300×300 Maxpooling 2: size = 2×2 stride = 2 FC: $n \times m \times m$, 10

[a] n, number of convolutional kernels; $h \times w$, dimension of each kernel. For FC layers, left-hand-side number is the input dimension, and right-hand-side number is the output dimension.

Structures of these two networks are shown in Table 1. In the pre-training phase, the dimension of each 2D CONV kernel is $h \times w$. The number of kernels is n (n equals eight in our architecture). The activation function is rectified linear unit (ReLU). Specifically, for models processing MNIST and Fashion MNIST, $h = w = 5$; for models processing Cifar10, $h = w = 9$. $m \times m$ is the original size of the input image (for example, $m = 32$ for Cifar10).

For the layout of the trained kernels on the 150×150 padding image, horizontal and vertical intervals between each kernel center are both 1.5 times the original input image size (e.g., 48 pixels for Cifar10) to avoid overlapping in the output plane. The transmission windows in corresponding positional filter sizes two times the input size in each dimension are spaced with an interval equal to three times the input size (e.g., window size equals 64×64, and interval equals 96 pixels for Cifar10), since the whole image is 300×300.

3. RESULTS AND DISSCUSSION

A. Fourier Convolution Tests

In Fourier convolution, the kernels are all Fourier-transformed into Fourier domain. We can derive the equivalent spatial

CONV kernel [or point spread function (PSF)] according to the modulation methods used in the Fourier plane. We substitute the classic convolutions with Fourier convolutions in our FCNN to make use of the $4f$ system to perform 2D Fourier transform and convolution. To validate the feasibility and benchmark the performance of our super-pixel complex-valued modulation method, we compare the single-channel and multi-channel FCNN results of different Fourier-domain-modulation methods in the following sections.

1. Single-Channel Fourier Convolution

In the simulated $4f$ system, a random image of the MNIST dataset is loaded onto the input plane, as shown in Fig. 3(a), while full light field complex amplitude distribution, AO distribution, PO distribution, and super-pixel-based complex-valued distribution of the Fourier transform of a single CONV kernel are loaded onto the Fourier plane, as shown in Fig. 3(b). It is referred to as single-channel convolution since there is only one output image in the output plane. Note that the pixel size of each input image here is enlarged four times from $6.3 \times 6.3\ \mu m^2$ to $25.6 \times 25.6\ \mu m^2$ to eliminate second-order diffraction while keeping the Fourier spectrum within the modulation area. Feature maps at the output image plane can be derived, as shown in Fig. 3(c), after the convolution is processed at the Fourier plane using different modulation methods. By the naked eye, it is hard to distinguish the differences among standard convolution, complex-field Fourier convolution, and super-pixel Fourier convolution. AO Fourier convolution also exhibits little difference with respect to the ideal convolution in this single-kernel situation. Noticeable blurring occurs only for PO Fourier convolution.

However, quantitative differences can be computed by the mean square errors (MSEs) of different methods with respect to standard spatial convolution. MSEs of different modulation methods with respect to standard spatially convolved feature maps [shown in Fig. 3(c)] are shown in Table 2. Average MSEs

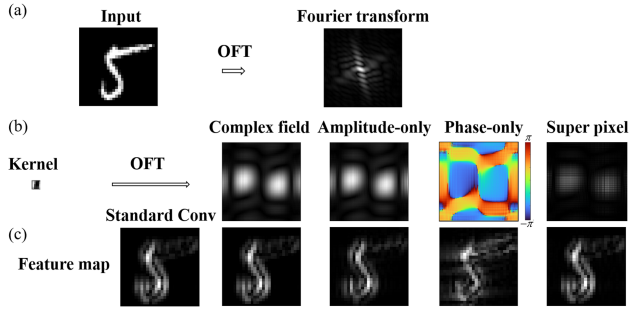

Fig. 3. Results of one-channel Fourier convolution tests with the convolutional kernel's Fourier transform modulated using different methods, including full complex field modulation, amplitude-only modulation, phase-only modulation, and the proposed super-pixel complex-valued modulation. (a) Input image and its Fourier transform. Random samples of MNIST are used as input images. (b) Fourier-plane modulations of a convolutional kernel using different methods. (c) Convolved feature maps of standard spatial convolution and different Fourier-plane modulation methods. Input size: 28×28, pixel size: $25.6 \times 25.6\ \mu m^2$; Fourier-plane size: 600×600, pixel size: $6.3 \times 6.3\ \mu m^2$; feature map size: 28×28, pixel size: $25.6 \times 25.6\ \mu m^2$.

Table 2. MSEs of Single-Channel Convolutional Feature Maps with Different Modulation Methods

Modulation Method	Mean Square Error (MSE)
Full-complex-field (ideal)	1.54×10^{-6}
Amplitude-only	6.30×10^{-3}
Phase-only	1.64×10^{-2}
Super-pixel	9.82×10^{-5}

are derived by averaging the MSEs of 100 random MNIST images. For one-channel convolutions, the loss of either amplitude information or phase information causes a large decrease in convolution accuracy. This is especially evident in the PO method, where the absence of amplitude modulation in the Fourier plane results in weaker spatial filtering capabilities and a larger MSE. On the other hand, the MSE of the proposed super-pixel method is similar to that of the ideal full-complex-field modulation, indicating a minimal loss in accuracy.

2. Multi-Channel Fourier Convolution

It is known that higher classification accuracy and robustness can be attained with multiple convolution kernels, which can be loaded sequentially in time. However, in the time-sequential modulation scheme, the repeatedly electronically driven updating of the weighted Fourier-transformed kernel is unavoidable, which leads to longer processing times and more power consumption. For small-sized input images and kernels, more than one CONV kernel can be tiled onto one kernel image, and multi-channels of convolved output images can be obtained simultaneously without cross talk. This is referred to as multi-channel convolution. In this method, we can transform the traditional time-sequential multi-kernel convolution into spatial multi-channel convolution by implementing complex-valued modulation in the Fourier plane, making full use of the parallelism of the free-space optical system and speeding up the CONV procedure by reducing the sequential steps. The same benefits of comparable classification accuracy and robustness can be achieved while taking much less time than the sequential multi-kernel method.

Assume the CONV kernel $W(x, y)$ is composed of an array of multiple CONV kernels. It can be expressed as

$$W(x, y) = \sum_{i=1}^{m} \sum_{k=1}^{n} W_{i,k}(x, y) * \delta(x - i\Delta x, y - k\Delta y), \quad \textbf{(1)}$$

where $W_{i,k}$ is the weighted kernel in column i, raw k; x, y are the spatial coordinates of the input plane; *denotes convolution; Δx, Δy are the spatial shifts of each kernel. The Fourier transform $F(f_x, f_y)$ of $W(x, y)$ can be written as

$$F(f_x, f_y) = \sum_{i=1}^{m} \sum_{k=1}^{n} F_{i,k}(f_x, f_y) \times \exp[-2\pi j(i f_x \Delta x + k f_y \Delta y)], \quad \textbf{(2)}$$

where f_x, f_y are the coordinates of the Fourier plane; $F_{i,k}$ is the Fourier transform of $W_{i,k}$. Generally, $F(f_x, f_y)$ in Eq. (2) is a complex-valued function, meaning that an ideal convolution can be achieved only through complex-valued modulation in

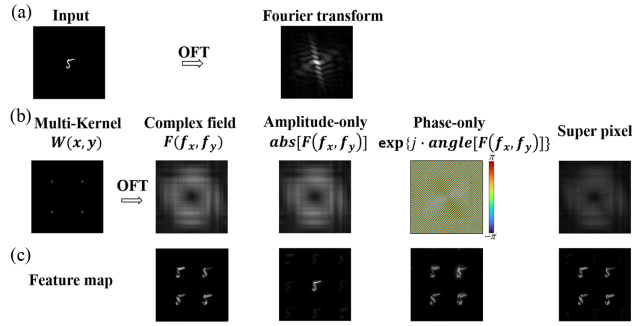

Fig. 4. Results of multi-channel Fourier convolution tests with different Fourier-plane modulation methods. (a) Input image and its Fourier transform. Input images are random samples of MNIST embedded into 600×600 paddings. (b) Fourier transforms of the convolutional kernels modulated in different methods. Example kernels here are classic Sobel operators detecting edges of different directions. They are tiled on one single image and Fourier-transformed. Specific Fourier-plane modulation method is noted on each modulation plane. (c) Convolved feature maps of standard spatial convolution and Fourier convolutions with different Fourier-plane modulation methods. The sizes of the input image, its Fourier transform at the Fourier plane, and the feature map are all 600×600, with a pixel size of $6.3 \times 6.3\ \mu m^2$.

the Fourier plane. However, commercial SLMs can perform PO or AO modulation, which are both insufficient for ideal convolution. The proposed super-pixel method provides a solution to this problem.

To be compatible with the angular spectrum (AS) method used in simulations of multi-channel convolutions, the input images should be pre-processed. The number of pixels of the input image is increased to 112×112, while the pixel size is kept at $6.3 \times 6.3\ \mu m^2$. Through these modifications, the physical size of the input image remains the same as in the case of single-channel convolution (pixel number: 28×28, pixel size: $25.6 \times 25.6\ \mu m^2$). Then the area of the input image is enlarged to 600×600 pixels with zeros padded in the empty surrounding area as shown in Fig. 4(a) to guarantee multi-channel output and ensure the adequate accuracy of the AS method [25].

To obtain the Fourier transform of the multi-kernel plane, multiple different CONV kernels are tiled onto a single plane and undergo the same preprocessing as the input images, including resizing and padding. The spacing between each kernel should be at least larger than the size of input image to prevent overlapping of different channels on the output plane. Afterwards, the Fourier transform of the entire image is calculated using diffraction-based Fourier transformation. Different Fourier-plane modulation methods are used to conduct multi-channel convolutions as shown in Fig. 4(b). The results are illustrated in Figs. 4(b) and 4(c).

From Fig. 4(b), it can be observed that the phase distribution of the Fourier transform of a multi-kernel plane is much more sophisticated than that of the single kernel in Fig. 3(b). Consequently, the loss of phase information will have a more severe impact on the convolution results. From Fig. 4(c), feature maps of the AO multi-channel Fourier convolution overlap at the center of the output plane, rendering it unsuitable for multi-kernel convolution. For PO modulation, errors are introduced in every output channel, causing the feature maps to be blurred. The MSEs of the multi-channel Fourier convolution with different modulation methods are shown in Table 3. From the derived feature maps and MSEs, the accuracy of PO modulation is similar to the case of single-channel convolution. Despite sacrificing half of the Fourier-plane resolution, the super-pixel method remains the most effective way to approximate full-complex-field modulation for use in multi-kernel convolution.

Table 3. MSEs of Multi-Channel Convolutional Feature Maps with Different Modulation Methods

Modulation Method	Mean Square Error (MSE)
Full-complex-field (ideal)	1.18×10^{-5}
Amplitude-only	Not applicable
Phase-only	1.60×10^{-2}
Super-pixel	2.53×10^{-4}

The above multi-channel convolution results can be understood from the analysis of Fourier-transform equations. In Eq. (2), $F_{i,k}(f_x, f_y)$ is the Fourier transform of the field distribution function of a kernel in the spatial domain, which can be approximated as the Fourier transform of a delta function under the approximation that the kernel is small enough compared with the whole padding plane. In this case, $F_{i,k}(f_x, f_y)$ is a constant corresponding to the amplitude of the kernel $W_{i,k}$ with no phase information. The phase information is all provided by the exponential factor related to the position of the kernel. Under these approximations, we can derive the distorted convolution kernel (or PSF) in the output image plane as follows if the amplitude/phase information is removed in its Fourier plane:

$$W_{\text{out-ao}}(x, y) \approx \sum_{i=1}^{m} \sum_{k=1}^{n} W_{i,k}(x, y) * \delta(x, y), \tag{3}$$

$$W_{\text{out-po}}(x, y) \approx \sum_{i=1}^{m} \sum_{k=1}^{n} \delta(x - i\Delta x, y - k\Delta y). \tag{4}$$

Here $W_{\text{out-ao}}$ is the inverse Fourier transform of F with AO information, and $W_{\text{out-po}}$ is the inverse Fourier transform of F with PO information. As shown in Eq. (3), for AO Fourier convolution, the CONV kernel tends to overlap at the center of the image plane, resulting in cross talk between output channels. Therefore, it is not suitable for multi-kernel convolutions. For PO Fourier convolution, although positional information is preserved, the weight information of different kernels is lost as illustrated in Eq. (4). As a result, the filtering effect is weakened, similar to the case of one-channel Fourier convolution in Section 3.A.1. Thus, the complex-valued modulation is even more important for multi-channel Fourier convolutions. Although the intensity of the feature map obtained by the super-pixel method is weaker (amplified by a factor of two for clearer illustration), the image quality is close to that obtained by using the ideal complex-valued modulation.

B. Multi-Channel FCNN Tests

Three classical datasets [MNIST, Fashion MNIST, and grayscale-CIFAR10, shown in Figs. 5(a)–5(c), respectively] are used to train and test our hybrid model. There are 50,000

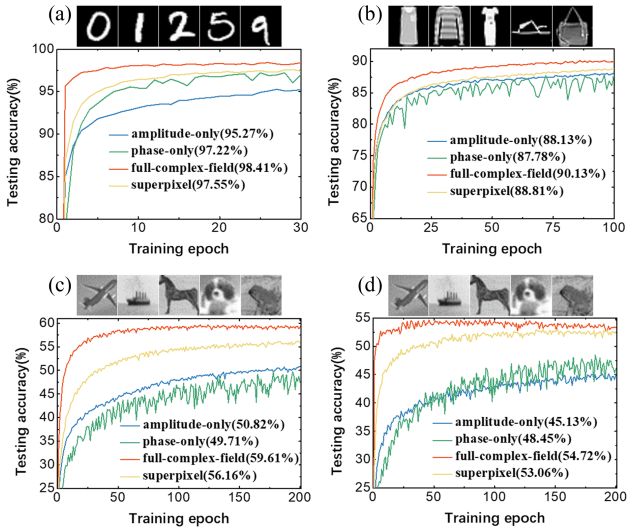

Fig. 5. Convergence plots of different models used for classification of MNIST, Fashion MNIST, and Cifar10, and the upside labels are representations of different datasets. The percentage in each inset denotes the peak accuracy of each model. (a) Convergence plots of eight-channel CNN, benchmarked with MNIST. (b) Convergence plots of eight-channel CNN, benchmarked with Fashion MNIST. (c) Convergence plots of eight-channel CNN, benchmarked with grayscale-Cifar10. (d) Convergence plots of four-channel CNN, benchmarked with grayscale-Cifar10.

training images and 10,000 testing images in each dataset. Before training, all images are preprocessed with the same procedure depicted in Section 3.A.2. In the training phase, training images are sent into the model and forward-propagated, and the derived logits vector of each image is compared with its label to calculate the loss function. SGD is applied to update the weights in the FC layer. In the testing phase, testing images are sent into the trained model and forward-propagated only to get logit vectors and compared with the labels to decide whether the predictions are correct. The test accuracy is worked out with the accurate predictions divided by the total amount of the testing images in one epoch. As shown in Fig. 5 and Table 4, while dealing with simple datasets such as MNIST, different modulation methods reach similar accuracy, because they do not need sophisticated feature extractions. For a slightly more complex dataset like Fashion MNIST, as shown in Fig. 5(b), the super-pixel method shows slightly better performance over the other two methods, as the accuracy of the super-pixel method can reach 88.8%. A more challenging dataset Cifar10 test further shows the advantages of complex-valued modulation since the super-pixel method can achieve blind test accuracy of 56.2%, while the AO and PO methods can reach only 50.8% and 49.7%, respectively. Results of four-channel architecture and eight-channel architecture are shown in Figs. 5(c) and 5(d), respectively, to prove that while only one-layer Fourier convolutions are performed, with the increasing number of kernels used, blind test accuracy improves within limited training epochs regardless of which modulation method is used.

Though in our architecture that uses the super-pixel in the Fourier plane, as the pixel size is 2× (in each dimension) larger than other methods, the corresponding object plane is half the region of the other methods, and the restricted object plane also restricts the number of CONV kernels loaded onto the same padding image. This restriction can be mitigated by applying SLM with higher resolutions.

Table 4. Peak Classification Accuracies of Eight-Kernel FCNN Model

Modulation Method	MNIST (%)	Fashion MNIST (%)	CIFAR10 (%)
Full-complex-field (ideal)	98.4	90.1	59.6
Amplitude-only	95.3	88.1	50.8
Phase-only	97.2	87.8	49.7
Super-pixel	97.6	88.8	56.2

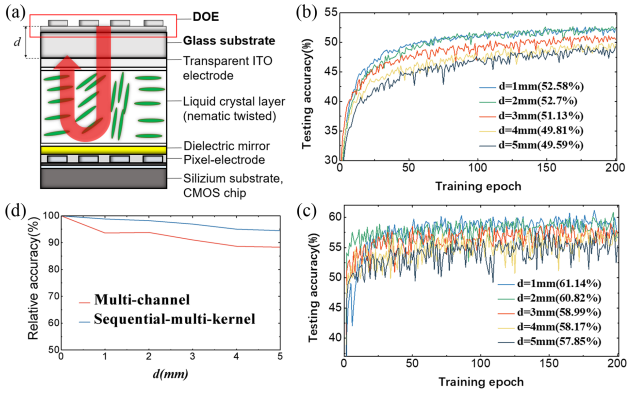

Fig. 6. (a) Sketch of structure between SLM and DOE considering the distance between them. (b) Convergence plots of multi-channel FCNN model with grayscale-Cifar10 of different gap sizes between the DOE and SLM. (c) Convergence plots of sequential-multi-kernel FCNN model grayscale-Cifar10 of different gap lengths. (d) Decay of accuracy with the gap length of two different methods.

C. Effect of the Gap between SLM and DOE

Generally, commercial liquid-crystal-based SLMs are packaged with a protective glass plate with a thickness of about 3 mm. The physical structure of a common twisted nematic LCOS [26] is shown in Fig. 6(a). In this scheme, the four-level DOE cannot be attached to its surface directly due to the gap between the DOE and the modulation plane, which means the diffraction between the DOE and the LCOS modulation unit cannot be ignored.

However, due to the inherent adaptive ability of the FC network, this diffraction effect can be moderated to some degree. The training plots of hybrid optical–electronic systems with different gap distances are shown in Figs. 6(b) and 6(c). The former indicates that the classification accuracy of multi-channel CNN decreases with an increased gap between the DOE and LCOS. When the gap is smaller than 4 mm, the fine-tuning capability of the FC layer will still have a blind test accuracy of approximately 51% with the multi-channel FCNN model, maintaining about 88% of its ideal value, as shown in Fig. 6(d). In this case, the sequential-multi-kernel FCNN model (in which kernels are loaded sequentially in time) seems to have better performance. For example, when the gap is as large as 5 mm, the sequential-multi-kernel method still has 57.85% classification accuracy, which is around 94% of its ideal accuracy

without the gap, and is about 10% higher than the multi-kernel method. That is because the tiled multi-kernel results in a larger size and consequently higher-frequency details in the Fourier plane, which leads to larger diffraction angles in the gap. For the sequential-multi-kernel method, there is only one kernel at a time; thus the corresponding simpler Fourier transform of the single kernel is not so sensitive to diffraction, as the adjacent super-pixels express similar values. But for the multi-channel method, the information loaded onto neighboring pixels is very different. The enlarged object field of the kernels corresponds to the increased details in the Fourier plane, leading to increased sensitivity to the variation of the gap between the DOE and SLM.

4. CONCLUSION

In summary, in this paper, a reconfigurable hybrid optical–electronic complex-value-modulated Fourier CNN with LCOS-based AO SLMs and a four-level DOE is proposed. The single-channel convolutions and multi-channel convolutions modulated with different methods are compared. The results show that the proposed super-pixel method, compared with the AO or PO method, is the best optical–electronic way to operate Fourier convolution, with the smallest MSE with respect to the standard spatial convolution. Evaluation of the hybrid FCNN models with different modulation methods is conducted using the benchmark datasets MNIST, Fashion MNIST, and Cifar10, in which the model with our proposed super-pixel method reaches state-of-the-art test accuracy of one-layer CNN models (97.55%, 88.81%, and 56.16%, respectively), which are all better than those obtained with AO or PO modulations, and close to the ideal complex-valued modulation method. The results suggest that quasi-complex-value modulation in the Fourier domain at the expense of lower Fourier-domain resolution shows better performance than the AO or PO method, especially for complex datasets. Furthermore, the added kernels in our one-layer Fourier convolution model improve the final blind test accuracy without increasing the processing time. The effect of the gap between the DOE and SLM has also been analyzed, which shows that the accuracy decreases with the increasing gap and increasing size of the tiled kernel. Therefore, this gap needs to be minimized as much as possible. The proposed hybrid optical–electric CNN computing system achieves complex-valued modulation in Fourier CNN with a single SLM, a four-level DOE, and a single optical $4f$ system, while achieving higher accuracy with better simplicity, experimental robustness, and adaptivity.

Funding. National Key Research and Development Program of China (2021YFB2801701); Huawei-ZJU Center for Innovation on Optical Computing.

Disclosures. The authors declare no conflict of interest.

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

[†]These authors contributed equally to this work.

REFERENCES

1. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Adv. Neural Inf. Process.Syst. **25**, 1097–1105 (2012).
2. L. Yann, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
4. H. Pan, "Learning convolutional neural networks in frequency domain," arXiv, arXiv:2204.06718 (2022).
5. G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," Nature **588**, 39–47 (2020).
6. C. Li, X. Zhang, J. Li, T. Fang, and X. Dong, "The challenges of modern computing and new opportunities for optics," PhotoniX **2**, 1 (2021).
7. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004–1008 (2018).
8. H. H. Zhu, J. Zou, H. Zhang, Y. Z. Shi, S. B. Luo, N. Wang, and A. Q. Liu, "Space-efficient optical computing with an integrated chip diffractive neural network," Nat. Commun. **13**, 1044 (2022).
9. C. S. Weaver and J. W. Goodman, "A technique for optically convolving two functions," Appl. Opt. **5**, 1248–1249 (1966).
10. S. Colburn, Y. Chu, E. Shilzerman, and A. Majumdar, "Optical frontend for a convolutional neural network," Appl. Opt. **58**, 3179–3186 (2019).
11. H. G. Chen, S. Jayasuriya, J. Yang, J. Stephen, S. Sivaramakrishnan, A. Veeraraghavan, and A. Molnar, "ASP vision: optically computing the first layer of convolutional neural networks using angle sensitive pixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 903–912.
12. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci. Rep. **8**, 12324 (2018).
13. M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, and V. J. Sorger, "Massively parallel amplitude-only Fourier neural network," Optica **7**, 1812–1819 (2020).
14. J. Xiang, S. Colburn, A. Majumdar, and E. Shlizerman, "Knowledge distillation circumvents nonlinearity for optical convolutional neural networks," Appl. Opt. **61**, 2173–2183 (2022).
15. T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, and Q. Dai, "Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit," Nat. Photonics **15**, 367–373 (2021).
16. C. K. Hsueh and A. A. Sawchuk, "Computer-generated double-phase holograms," Appl. Opt. **17**, 3874–3883 (1978).
17. L. G. Neto, D. Roberge, and Y. Sheng, "Full-range, continuous, complex modulation by the use of two coupled-mode liquid-crystal televisions," Appl. Opt. **35**, 4567–4576 (1996).
18. E. Ulusoy, L. Onural, and H. M. Ozaktas, "Full-complex amplitude modulation with binary spatial light modulators," J. Opt. Soc. Am. A **28**, 2310–2321 (2011).
19. S. A. Goorden, J. Bertolotti, and A. P. Mosk, "Superpixel-based spatial amplitude and phase modulation using a digital micromirror device," Opt. Express **22**, 17999–18009 (2014).
20. S. Jiao, D. Zhang, C. Zhang, Y. Gao, T. Lei, and X. Yuan, "Complex-amplitude holographic projection with a digital micromirror device (DMD) and error diffusion algorithm," IEEE J. Sel. Top. Quantum Electron. **26**, 2800108 (2020).
21. H. M. Ozaktas and H. Urey, "Space-bandwidth product of conventional Fourier transforming systems," Opt. Commun. **104**, 29–31 (1993).
22. P. Birch, R. Young, D. Budgett, and C. Chatwin, "Dynamic complex wave-front modulation with an analog spatial light modulator," Opt. Lett. **26**, 920–922 (2001).
23. E. G. Van Putten, I. M. Vellekoop, and A. P. Mosk, "Spatial amplitude and phase modulation using commercial twisted nematic LCDs," Appl. Opt. **47**, 2076–2081 (2008).

24. L. Fan, B. Liu, X. Long, C. Li, X. Dong, J. Cheng, and J. J. He, "Optical convolutional neural network based on an amplitude modulation SLM and a 4-level phase plate," Proc. SPIE **11898**, 11898O (2021).
25. K. Matsushima and T. Shimobaba, "Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields," Opt. Express **17**, 19662–19673 (2009).
26. Z. Zhang, Z. You, and D. Chu, "Fundamentals of phase-only liquid crystal on silicon (LCOS) devices," Light Sci. Appl. **3**, e213 (2014).