Research Article | Vol. 31, No. 4 / 13 Feb 2023 / Optics Express 6453

Optics EXPRESS

Co-designed metaoptoelectronic deep learning

CARLOS MAURICIO VILLEGAS BURGOS,[1] PEI XIONG,[1] LIANGYU QIU,[1] YUHAO ZHU,[2,3] AND A. NICKOLAS VAMIVAKAS[1,*]

[1] University of Rochester, Institute of Optics, 275 Hutchison Road, Rochester, NY 14627, USA
[2] University of Rochester, Department of Computer Science, 2513 Wegmans Hall, Rochester, NY 14627, USA
[3] yzhu@rochester.edu
[*] nick.vamivakas@rochester.edu

Abstract: A metaoptical system is co-designed with electronic hardware to implement deep learning image recognition. The optical convolution block includes a reflective metasurface to perform one layer of a deep neural network. The optical and digital components are jointly optimized to perform an image classification task attaining 65% accuracy, which is close to the 66% accuracy of a fully-digital network where the optical block is replaced by a digital convolution layer.

1. Introduction

The set of algorithms known as deep learning (also known as artificial neural networks or deep neural networks) have become ubiquitous in many technological fields because of their applicability in a wide array of automated tasks [1,2]. Despite having been introduced many decades ago, they have gained notable traction in recent years due to technology improvements, such as more powerful electronic hardware allowing for more efficient computations and larger storage sizes [2]. Some of the most common applications of deep learning algorithms include computer vision [3,4] and language processing [5–7]. Optical systems that implement deep learning were proposed several decades ago [8]. However, the recent rise of deep learning has revived interest in optical hardware implementations. The main motivations are the potential savings in computation and energy costs attained by performing mathematical operations on optical hardware instead of electronic hardware. Demonstrations include implementing existing neural network architectures, such as a multi-layer perceptron [9,10], a convolutional neural network [11–18], and recurrent neural networks [19,20], while others implement unique architectures enabled by optics [21–29] or use the optical hardware for data transfer [30] or logic operations [31]. These are in addition to applications relying on image deconvolution common in both astronomy [32] and microscopy [33].

In addition to optical systems that intend to delegate computational overhead into the optics, there are works that propose optical systems that are designed using deep learning algorithms to perform domain-specific tasks, in what has been termed as deep optics [34,35]. Deep optics systems have found applications in tasks that are relevant for computer vision such as depth estimation [36] and acquisition of high-dynamic-range images [37]. Other examples of recent works on this field include the design of imaging systems with large depth-of-field [38], end-to-end joint design of the lenses in an imaging system [39], design of multi-channel imaging systems for fast acquisition of depth information in microscopy systems [40], and the joint design of nano-optics imagers and image reconstruction algorithms for a high-quality and ultra-compact computational imaging system [41].

In this work we extend the field of co-designed deep optics to include metasurface optical components and provide an experimental proof-of-concept for the design framework introduced in [18]. Metasurfaces enable flexible manipulation of light's amplitude, phase and polarization [42] and make possible the engineering of an optical system's point spread function. The metasurface

#479038 https://doi.org/10.1364/OE.479038
Journal © 2023 Received 26 Oct 2022; revised 9 Dec 2022; accepted 13 Dec 2022; published 8 Feb 2023

Optics EXPRESS

used in this work is a plasmon gradient metasurface; a type of reflective metasurface [43]. Some recent works have used metasurfaces to implement deep learning concepts [14,22,25,29,31,41]. However, all of these works only use transmissive metasurfaces, which has left open opportunities to demonstrate the usage of reflective metasurfaces. As such, this work is the first instance of a deep optics system where a reflective metasurface is used and co-designed with digital neural network layers. This is significant, as reflective elements allow light to follow paths that are not restricted to just a straight line, which is useful when there are geometric constraints on the optical system. Additionally, our design pipeline can be generally used in other deep optics systems that incorporate digital neural network layers and that perform mathematical operations using a phase-modulating optical element, such as a metasurface. Furthermore, our work is the first one to perform joint training of the parameters of a reflective metasurface along with the parameters of the digital components in a hybrid system that combines them to perform an image classification task.

2. System design

A natural fit for implementing the convolution mathematical operation using optical hardware is to use a system based on free-space propagation, as both far-field propagation and propagation through lenses perform Fourier transform operations of the optical field at no additional energy cost [44]. Because of this, the intensity distribution of the images that are captured by a linear optical system with incoherent illumination can be expressed in terms of a convolution operation, that is:

$$I_{\text{out}}(u, v) = I_{\text{in}}(\xi, \eta) * \text{PSF}(\xi, \eta), \tag{1}$$

where $*$ denotes the convolution operation, I_{in} is the intensity distribution in the optical system's input plane, and PSF is the optical system's point-spread function (PSF), which can be viewed as the image captured by the optical system when it is imaging a point source [44].

The concept for this work is shown in Fig. 1. As previously stated, the goal is to set up and test an optical system that can implement the mathematical operations performed by a digital convolutional layer and reproduce the outputs that the latter would yield. This can be achieved if the convolution kernels of that digital convolutional layer are encoded in the optical system's PSF. The optical system's PSF can be engineered by adjusting the phase modulation profile imparted by a metasurface.

The output numerical array yielded by this optical system will differ slightly from the one that would have been yielded by the original digital convolutional layer. However, since that convolutional layer doesn't exist in a vacuum and is instead part of an artificial neural network whose purpose is to perform a task (such as object detection or image classification), we are more interested in the performance that the rest of the network (referred to as the suffix layers moving forward) would have when it takes the output of the optical convolution system as an input instead of taking the output of the original digital convolutional layer. Because of this, the process for testing the metasurface-based optical convolution system is designed around measuring the performance of the joint hybrid system that is comprised by the optical convolution block and the digital suffix layers.

We follow a design pipeline, which was proposed in our previous work [18], where a sequence of steps are followed in order to train the parameters of both the optical convolution block and the digital suffix layers. This pipeline can be summarized as follows. First, a fully-digital neural network is trained to classify images from the CIFAR-10 dataset, created by the Canadian Institute for Advanced Research (from which it gets its name) [45]. Then, the first convolutional layer of this network is replaced by a parametrized model of the optical convolution block, and this model's parameters are co-trained along with those of the digital suffix layers. The optical model's trained parameters are then used to fabricate the metasurface that is present in the physical optical convolution block. Finally, the parameters of the suffix layers need to

Optics EXPRESS

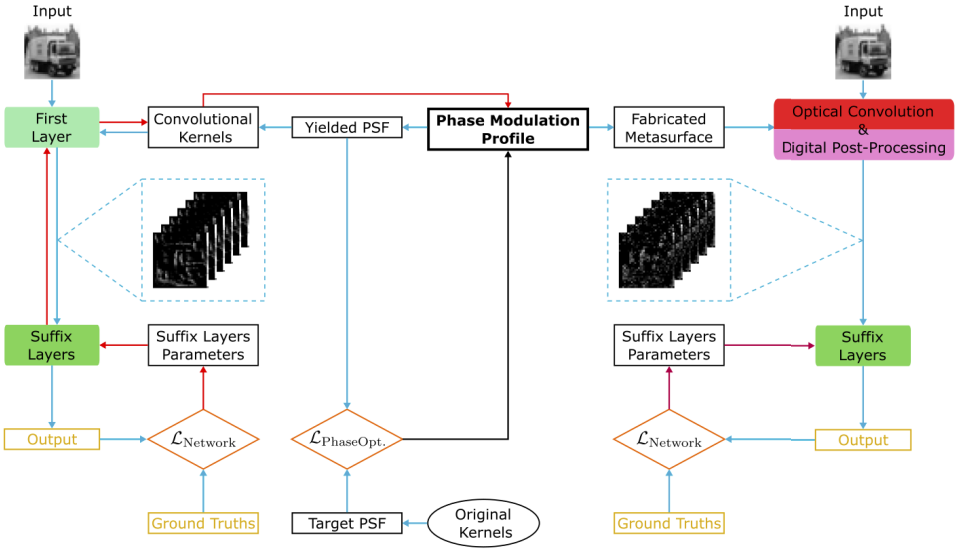

Fig. 1. Our optical convolution system aims to reproduce the same result as a digital convolutional layer. The response of the optical system is characterized by its point-spread function (PSF), which is controlled by the phase modulation profile of a metasurface. This phase modulation profile is first trained so that the optical system's yielded PSF encodes the convolution kernels of the digital convolution layer that is being replaced by it. Then, it is co-trained along with the parameters of the suffix digital layers to optimize the system's performance on an image classification task. After that, the co-trained phase profile is used to fabricate the optical system's metasurface, and the hybrid system is able to carry out the image classification task. Finally, the parameters of the hybrid system's digital suffix layers are fine-tuned to improve its classification performance.

be fine-tuned to account for the differences between the outputs of the optical system and the original digital layer that was replaced by the former.

The key elements of this pipeline are the algorithms that are used to encode the convolution kernels into the optical system's PSF, as well as jointly training the parameters of the optical convolution block with those of the digital suffix layers. The former is done with a phase optimization algorithm, where the phase modulation profile of the metasurface is optimized via gradient descent so that its yielded PSF approaches a target PSF that encodes the digital layer's convolution kernels. Meanwhile, the latter is accomplished by performing backpropagation training (which is another form of gradient descent optimization) on the differentiable model that results from parametrizing the network's first layer's convolution kernels in terms of the phase modulation profile of the optical convolution block's metasurface. In order to perform both of the aforementioned tasks, it is necessary to compute the gradient $\nabla_\Phi \mathcal{L}$ of a task-related loss function $\mathcal{L}$ with respect to the phase modulation profile Φ. These processes are illustrated in Fig. 1.

In the case of the phase optimization task, the loss function $\mathcal{L}_{\text{PhaseOpt.}}$ that is to be minimized via gradient descent is given by:

$$\mathcal{L}_{\text{PhaseOpt.}}(\Phi;\, \text{PSF}_{\text{target}}) = ||\,(\text{PSF}_{\text{target}}/\text{ATT}) - |\mathcal{F}_{\text{2D}}^{-1}\left\{e^{i\Phi}\right\}|^2\,||_F^2, \tag{2}$$

where $\text{PSF}_{\text{target}}$ is the target PSF that encodes the original digital layer's convolution kernels, ATT is an attenuation profile that affects the PSF yielded by the metasurface, $\mathcal{F}_{\text{2D}}^{-1}$ denotes an invese 2D Fourier transform, and $||\cdot||_F$ denotes the Frobenius norm. More technical details regarding the attenuation profile ATT can be found in the Supplement 1. The gradient of $\mathcal{L}_{\text{PhaseOpt.}}$ with

Optics EXPRESS

respect to Φ is given by:

$$\nabla_{\Phi}\mathcal{L}_{\text{PhaseOpt.}} = -4\text{Im}\left[e^{-i\Phi}\mathcal{F}_{\text{2D}}\left\{\left(Y - \left|\mathcal{F}_{\text{2D}}^{-1}\left\{e^{i\Phi}\right\}\right|^2\right)\mathcal{F}_{\text{2D}}^{-1}\left\{e^{i\Phi}\right\}\right\}\right], \tag{3}$$

where $Y = (\text{PSF}_{\text{target}}/\text{ATT})$, and Im denotes the function that returns the imaginary part of an array of complex numbers.

In the case of the co-training task, where the phase modulation profile is jointly trained with the system's digital layers, the loss function $\mathcal{L}_{\text{Network}}$ to be minimized is the cross-entropy loss function (also known as log-loss). This is the same loss function that is minimized when a fully-digital network is trained to perform its classification task. The log-loss function measures the "confidence" the system has in assigning the correct class label (ground truth) to each image that it is classifying, as the value of the log-loss function is lower if the predicted probabilities assigned by the system to the ground truth labels are higher [2]. When the original fully-digital network is trained, the convolution kernels W of the first layer are iteratively updated by computing $\nabla_W\mathcal{L}_{\text{Network}}$. However, when these kernels are parametrized in terms of Φ, Φ needs to be updated by computing the gradient $\nabla_{\Phi}\mathcal{L}_{\text{Network}}$ instead. The latter gradient can be computed in terms of the former as:

$$\nabla_{\Phi}\mathcal{L}_{\text{Network}} = 2\text{Im}\left[e^{-i\Phi}\mathcal{F}_{\text{2D}}\left\{F\left(\nabla_W\mathcal{L}_{\text{Network}}\right)\mathcal{F}_{\text{2D}}^{-1}\left\{e^{i\Phi}\right\}\right\}\right], \tag{4}$$

where $F\left(\nabla_W\mathcal{L}_{\text{Network}}\right) = \nabla_{\mathbf{PSF}}\mathcal{L}_{\text{Network}}$ is the gradient of $\mathcal{L}_{\text{Network}}$ with respect to the PSF yielded by Φ, and can be obtained by re-arranging the elements of $\nabla_W\mathcal{L}_{\text{Network}}$. More details about the relationship between $\nabla_W\mathcal{L}_{\text{Network}}$ and $\nabla_{\mathbf{PSF}}\mathcal{L}_{\text{Network}}$, as well as rigorous derivations of Eq. (3) and Eq. (4) can be found in our previous work [18].

3. Results

The metasurface's phase modulation profile was obtained after finishing the co-training steps of the design pipeline, where it was optimized alongside the parameters of the digital suffix layers in order to perform an image classification task on images from the CIFAR-10 dataset. Before fabricating the metasurface, a fully-digital neural network was built where the first layer's convolution kernels were derived from the co-trained phase modulation profile, and the suffix layer's parameters were set to those obtained at the end of the co-training steps. The full network had a classification performance of 86% accuracy and a log-loss of 0.50 on the test set of the CIFAR-10 dataset.

After fabricating the metasurface, it was placed in the optical system to perform optical convolution. Images from the test set were projected into the optical convolution block, and the yielded captures were saved into a computer's hard drive. An example capture can be found in Fig. 2(a)). The captures went through digital post-processing to produce tensors that were used as inputs to the digital suffix layers, in order to get a set of classification predictions for every input image. More details on this experiment can be found in the Methods section. The initial classification performance of this hybrid system was 11% accuracy and 4.48 log-loss, before any fine-tuning of the digital suffix layers was done. However, after the parameters of the suffix layers were fine-tuned to account for the differences between the inputs coming from the original first digital layer and those coming from the optical convolution block, the hybrid system's performance increased to 65% accuracy and 1.09 log-loss.

This performance was benchmarked against a fully-digital network that was constructed by setting the parameters of its suffix layers to have the values they had after the final fine-tuning step, while the convolution kernels of its first layer were set to have the values obtained from the metasurface's phase modulation profile. A diagram illustrating the construction of this benchmark digital network can be found in Fig. 2(b)). By construction, the only difference between the benchmark network and the hybrid system is that the digital first layer is being

Optics EXPRESS

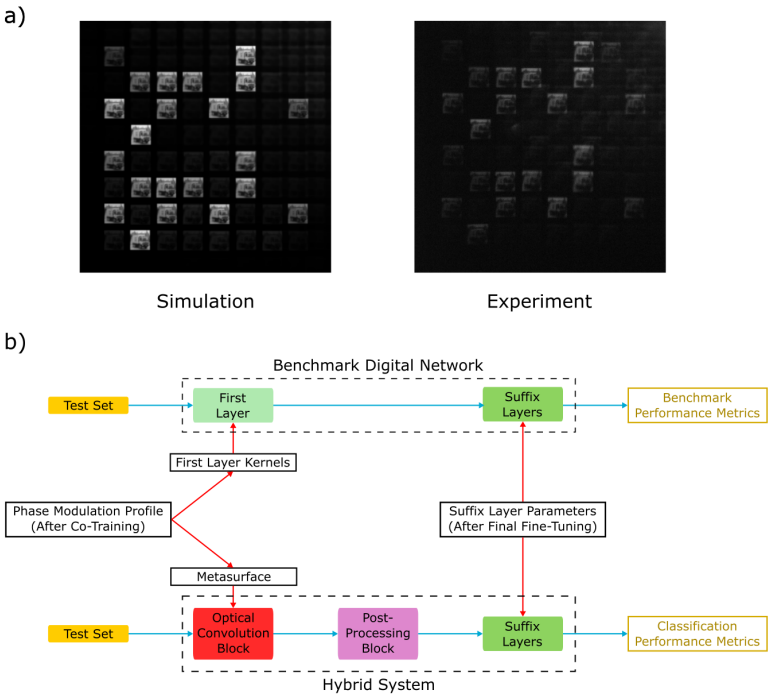

Fig. 2. a) Example capture: simulation vs experiment. b) Construction of the benchmark digital neural network.

replaced by the optical convolution block in the latter system, and both systems have the same parameters in their suffix layers. The benchmark network has a performance with an accuracy of 66% and a log-loss of 1.11. Given this, it can be seen that the performance remains almost the same when the digital first layer is replaced by the optical convolution block in the hybrid system.

Despite the similar classification performance of the hybrid system and the digital benchmark network, the predictions yielded by each system for a given individual input tend to be different. This is to be expected, due to the qualitative differences that exist between the outputs yielded by the digital and physical versions of the network's first layer, despite the latter being designed to reproduce the outputs of the former. An example qualitative comparison between both can be found in Fig. 3. In the final step of the optimization pipeline, the digital suffix layers were fine-tuned so that they could improve their classification performance when receiving inputs coming from the optical convolution block, but it is important to note that this fine-tuning process would not necessarily translate into having the suffix layers yield the same response when receiving inputs coming from the optical convolution block as when receiving inputs coming from a digital layer. That is, the differences between both types of inputs gives rise to differences between the corresponding outputs yielded by the same suffix layers. As such, there's differences between the responses that the hybrid system and the fully-digital benchmark network will have for a given input image. We quantify these differences by comparing the predictions given by both systems when receiving pictures from the test set of the CIFAR-10 dataset as inputs. Out of the 10000 pictures in the test set, both systems coincided in their predicted classes for 5482 of these pictures. Out of these 5482 overlapping predictions, 4918 were correct predictions (predicted class matching ground truth class). However, beyond those overlaps, the hybrid system got 1601 correct predictions that the digital benchmark network got incorrect, while the latter got other 1699 correct predictions that the former got incorrect, and there were 1782 pictures

Optics EXPRESS

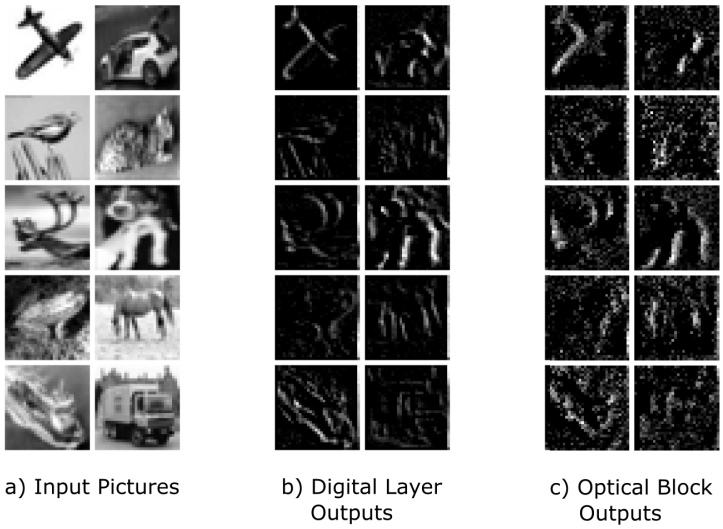

Fig. 3. a) Sample inputs, one for each of the ten possible classes in the dataset. b) Corresponding outputs yielded by the digital first layer. c) Corresponding outputs yielded by the optical convolution block (after post-processing). In both b) and c), only one channel of the three-dimensional output tensors is shown.

where both systems gave incorrect predictions (though they coincided in the same wrong answer on only 564 out of these pictures). In total, the hybrid system got 6519 correct predictions and the digital benchmark network got 6617 correct predictions, giving rise to the 65% accuracy for the former and 66% accuracy for the latter that were reported above. We do further quantitative comparisons by computing the structural similarity index measure (SSIM) and root-mean-square difference (RMSE) between the outputs coming from each version of the network's first layer. Before doing so, the output tensors associated with each picture in the dataset were normalized to have a maximum value of 1, to have both versions of each tensor have the same dynamic range

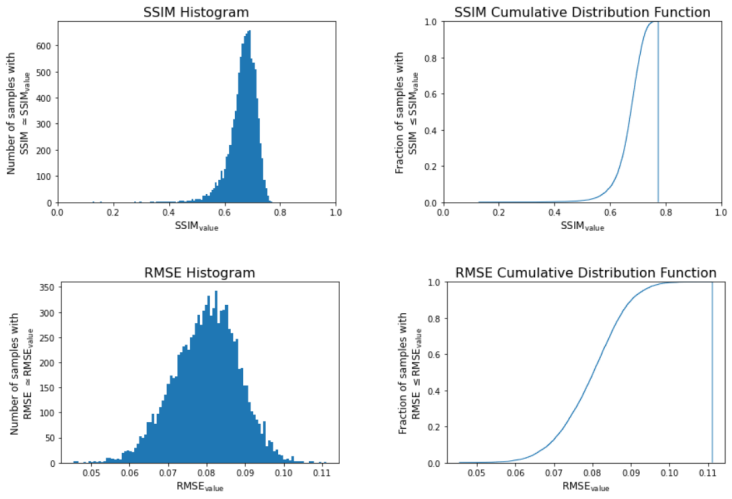

Fig. 4. Histograms and cumulative distribution functions of the SSIM and RMSE between the outputs yielded by the physical and digital versions of the network's first layer.

Optics EXPRESS

at the time of computing the SSIM and RMSE. In Fig. 4, we show histograms and cumulative distribution functions of the values that the SSIM and RMSE metrics have when comparing the outputs yielded by both versions of the network's first layer when given each picture from the CIFAR-10 test set as their input. The average SSIM is 0.66, and the average RMSE is 0.0798, which is 7.98% of the dynamic range of the compared tensors.

4. Methods

4.1. Optical system layout

A diagram illustrating the layout of the optical convolution block is shown in Fig. 5. In this system, 633 nm light from a HeNe laser is imaged out-of-focus into the surface of a rotating glass diffuser, and the diffused light is collimated and sent into the surface of a DLP LightCrafter 6500 digital micro-mirror display (DMD). The light incident on the DMD has become spatially incoherent after going through the rotating diffuser, and it overfills the array of display pixels that are used to project pictures into the optical convolution block. The display plane and lenses L3 and L4 are arranged in a 4f system, which is a configuration used to perform optical convolution [44]. That is, the display plane is placed in the front focal plane of lens L3, and the metasurface is placed on the back focal plane of this lens, which coincides with the front focal plane of lens L4 and is the Fourier plane of the 4f system. Both lenses L3 and L4 have a focal length of 125 mm. Since the metasurface is reflective, a pellicule beam-splitter is placed between L3 and the metasurface, so that light reflected by the latter can be pathed into lens L4. Additionally, since light incident on the metasurface needs to have vertical polarization so that it can impart the intended phase values on the reflected light, a linear polarizer is also added between lens L3 and the beam-splitter.

An image that consists of the convolution between the picture projected on the display and the PSF yielded by the metasurface is formed on the output plane of this first 4f system, which is the back focal plane of lens L4. This plane coincides with the input plane of a second 4f system composed by lenses L5 and L6. The metasurface and the beam-splitter are tilted at angles that allow the portion of interest in the 4f system's output image to be incident at the center of the clear aperture of mirror M5, and the tilt angle of this mirror is adjusted so that reflected light can propagate along the direction of the line that connects the centers of lenses L5 and L6. Lens L5 has a focal length of 150 mm and lens L6 has a focal length of 15 mm, so this 4f system has a magnification of $-\frac{1}{10}$. The goal of this second 4f system is to have the portion of interest of the image that is formed at its input plane shrunken down in size so that it can fit into the Basler acA1600-20um camera's sensor. The camera's capture contains an array of sub-images that are equal to the individual convolutions between the picture projected on the DMD and the convolution kernels that have been encoded on the system's PSF by the metasurface's phase profile.

4.2. Capture acquisition process

Once the alignment of the components in the optical system is completed, the capture acquisition process is performed to obtain captures containing the system outputs associated with each of the pictures in the CIFAR-10 dataset. Captures associated with both the training set and the test set are obtained on the same experiment session, so that they are acquired under the same conditions. In order to sequentially project the hundreds of pictures contained in each set, video files for both sets are created. In these videos, each frame contains a different picture and the video is played at 10 frames per second. The computer that controls the DMD and the camera is set-up to play these videos on the DMD, which is configured as a second display. Meanwhile, the camera is set to acquire captures at 20 frames per second, which is twice the frequency at which pictures in the video are changed. This is done to avoid getting captures of a transitory state of the DMD (when

Optics EXPRESS

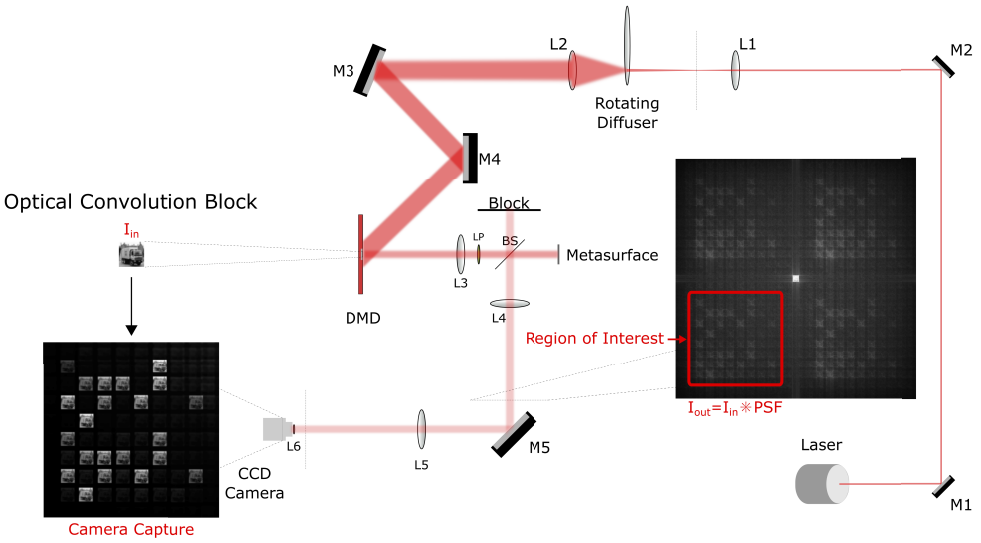

Fig. 5. Optical convolution block. Light from a laser is passed through some optical elements (labeled with letters L for lenses and M for mirrors) to make it become spatially incoherent before it illuminates the surface of a digital micro-mirror display (DMD). The illuminated DMD projects images, and light coming from it passes through lens L3, a linear polarizer (LP), and a pellicule beam-splitter (BS) before being incident on the metasurface. Light reflected by the metasurface is then reflected by the beam-splitter into lens L4. An image is formed in the back focal plane of lens L4, which consists of the convolution between the input picture projected by the DMD and the point spread function of the optical system, which is controlled by the metasurface's phase modulation profile. The region of interest containing the result of the optical convolution is imaged into a CCD camera's sensor using mirror M5 and lenses L5 and L6.

it switches from a picture to another). However, the camera is also set up to save only every second capture it acquires to the computer's hard disk. That way, there is one saved capture for every picture in the dataset (for both the training set and the test set).

4.3. *Post-processing steps*

After these captures have been acquired, it is necessary to perform some digital post-processing steps on them in order to obtain a set of tensors (three-dimensional numerical arrays) that reproduce the outputs that the original digital first layer would have yielded. These tensors will serve as the input to the network's digital suffix layers in order to complete the image classification task. The post-processing steps consist on extracting the patches of the capture that contain the sub-images (which represent the results of individual convolutions between the input picture and each convolution sub-kernel), as well as additional correction steps on the extracted sub-images, such as resizing and brightness compensation (by multiplying the sub-image in question by a scalar).

As explained in detail in the Supplement 1 document, each of the convolution kernels W contained in the optical system's PSF are split into two sub-kernels each; one sub-kernel containing the positive values of W and the other one containing the (absolute value of) the negative values of W. An image captured by the sensor contains an array of sub-images that are equal to the results of convolving the corresponding input picture I_{in} with the sub-kernels contained in the PSF, which encode the convolution kernels W of the replaced first digital convolutional layer.

Optics EXPRESS

Subtracting each sub-image with its corresponding pair during post-processing will yield a result that is equivalent to convolving I_{in} with W.

After all the sub-images contained in the capture are extracted and corrected, they are subtracted with their corresponding pair in order to obtain the results of the convolutions between the input picture and all the convolution kernels. Then, these convolution results are stacked in a 3-D tensor where the first two dimensions are width and height, and the third dimension is the channel index; this is the same format as the outputs of a digital convolutional layer. Finally, a non-linear function is applied element-wise on the resulting tensor, to obtain an output that nearly reproduces the one that would have been yielded by the original digital convolutional layer. The non-linear function that is used both in the digital convolutional layers and this final post-processing step is the ReLU function, which maps non-negative numbers to themselves and negative numbers to zero. An example of a capture and the result of the post-processing steps applied to it are shown on Fig. 3, along with a comparison to the output of the original digital convolutional layer.

After the post-processing steps are performed on all the saved captures, the results are stacked in a 4-D tensor dataset, where the first dimension is the capture/picture index, the second and third dimensions are width and height, and the fourth dimension is the channel index. There is one such tensor constructed from the captures from the test set, and one constructed from the captures of the training set. The former is used to measure the hybrid system's performance in the image classification task, while the latter is used to fine-tune the parameters of the system's digital suffix layers.

5. Discussion and conclusion

In this proof-of-concept work we implemented a metasurface-based optical system that can perform the same mathematical operations as a convolutional layer from an artificial neural network, delegating some of the computation overhead from the electronic hardware into the optical system. While the performance of the hybrid system in this work is lower than that of the original fully-digital network, it has a performance that is on par to that of the benchmark fully-digital network that has the same values on the parameters of its suffix layers, despite the differences between the outputs yielded by the optical and digital versions of the system's first layer.

Since this is a proof-of-concept, our main goal was not to have an original fully-digital network with state-of-art performance, nor have the system maintain such performance after replacing the network's first layer with the metasurface-based optical convolution block. Rather, the goal of this work was to experimentally demonstrate the process that is followed to design and fabricate a metasurface that can be used as part of a system that can perform a computer vision task (like image classification in this case). Additionally, since the metasurface in question was reflective, this proof-on-concept demonstrates that optical systems that intend to use a metasurface-based approach to perform deep learning computations do not need to be constrained to only use transmissive elements, allowing them to have more flexibility in their design.

The advantages in terms of reducing computation costs and latency by performing deep learning operations on optical hardware have been discussed in other works more extensively. An additional advantage that a system like this can have is the ability to introduce privacy preservation into the optical systems where metasurfaces can be integrated due to their compact and light-weight form factor, such as cell-phone cameras and augmented reality gear. This privacy preservation element can be implemented by modifying the convolution kernels of the replaced convolution layer or by introducing aberrations into the optics, in such a way that the captures on the sensor are no longer human-perceptible, while the post-processing steps and the digital suffix layers ran in the electronic portion of the system ensure that the computer vision tasks are still carried out with high performance. Such approaches would have the benefits of simultaneously preserving privacy and running computations on the optical hardware. However,

Optics EXPRESS

while such possibilities open up interesting avenues for future research, they are out of the scope of this work.

Funding. University of Rochester University Research Award.

Acknowledgments. C. Villegas Burgos would like to thank the Science and Technology Council of Mexico (Consejo Nacional de Ciencia y Tecnología, CONACYT) for the financial support provided by the fellowship they granted him, as well as S. Dadrasmarani and P. Sharma for their technical assistance in the laboratory during the early stages of the project. The authors thank the University of Rochester for supporting this work with a University Research Award. The authors would also like to thank F. Cheng for his contribution in creating the library that contains the field-modulation response yielded by nano-tokens with different geometries.

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.

Supplemental document. See Supplement 1 for supporting content.

References

1. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
2. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.
3. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 779–788.
4. L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr, "Fully-convolutional siamese networks for object tracking," in *European conference on computer vision*, (Springer, 2016), pp. 850–865.
5. T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, (2010), pp. 1045–1048.
6. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, (2013), pp. 3111–3119.
7. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, (2014), pp. 3104–3112.
8. C. Denz, *Optical Neural Networks* (Vieweg Teubner Verlag, 1998).
9. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
10. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X **9**(2), 021032 (2019).
11. 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**(1), 12324 (2018).
12. M. Miscuglio, Z. Hu, S. Li, J. K. George, R. Capanna, H. Dalir, P. M. Bardet, P. Gupta, and V. J. Sorger, "Massively parallel amplitude-only fourier neural network," Optica **7**(12), 1812–1819 (2020).
13. S. Colburn, Y. Chu, E. Shilzerman, and A. Majumdar, "Optical frontend for a convolutional neural network," Appl. Opt. **58**(12), 3179–3186 (2019).
14. C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, "Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network," Nat. Commun. **12**(1), 96 (2021).
15. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**(7840), 44–51 (2021).
16. Z. Gu, Y. Gao, and X. Liu, "Optronic convolutional neural networks of multi-layers with different functions executed in optics for image classification," Opt. Express **29**(4), 5877–5889 (2021).
17. A. Ryou, J. Whitehead, M. Zhelyeznyakov, P. Anderson, C. Keskin, M. Bajcsy, and A. Majumdar, "Free-space optical neural network based on thermal atomic nonlinearity," Photonics Res. **9**(4), B128–B134 (2021).
18. C. M. V. Burgos, T. Yang, Y. Zhu, and A. N. Vamivakas, "Design framework for metasurface optics-based convolutional neural networks," Appl. Opt. **60**(15), 4356–4365 (2021).
19. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**(6), 756–760 (2018).
20. G. Mourgias-Alexandris, G. Dabos, N. Passalis, A. Totovic, A. Tefas, and N. Pleros, "All-optical wdm recurrent neural networks with gating," IEEE J. Sel. Top. Quantum Electron. **26**(5), 1–7 (2020).
21. 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**(6406), 1004–1008 (2018).
22. Z. Wu, M. Zhou, E. Khoram, B. Liu, and Z. Yu, "Neuromorphic metasurface," Photonics Res. **8**(1), 46–50 (2020).
23. H. Zhang, M. Gu, X. D. Jiang, J. Thompson, H. Cai, S. Paesani, R. Santagati, A. Laing, Y. Zhang, M. H. Yung, Y. Z. Shi, F. K. Muhammad, G. Q. Lo, X. S. Luo, B. Dong, D. L. Kwong, L. C. Kwek, and A. Q. Liu, "An optical neural chip for implementing complex-valued neural network," Nat. Commun. **12**(1), 457 (2021).

Optics EXPRESS

24. 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**(5), 367–373 (2021).
25. X. Zhang, Y. Zhou, H. Zheng, A. E. Linares, F. C. Ugwu, D. Li, H.-B. Sun, B. Bai, and J. G. Valentine, "Reconfigurable metasurface for image processing," Nano Lett. **21**(20), 8715–8722 (2021).
26. T. Wang, S.-Y. Ma, L. G. Wright, T. Onodera, B. C. Richard, and P. L. McMahon, "An optical neural network using less than 1 photon per multiplication," Nat. Commun. **13**(1), 123 (2022).
27. L. G. Wright, T. Onodera, M. M. Stein, T. Wang, D. T. Schachter, Z. Hu, and P. L. McMahon, "Deep physical neural networks trained with backpropagation," Nature **601**(7894), 549–555 (2022).
28. Y. Li, R. Chen, B. Sensale-Rodriguez, W. Gao, and C. Yu, "Real-time multi-task diffractive deep neural networks via hardware-software co-design," Sci. Rep. **11**(1), 11013 (2021).
29. C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li, and T. J. Cui, "A programmable diffractive deep neural network based on a digital-coding metasurface array," Nat. Electron. **5**(2), 113–122 (2022).
30. L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Freely scalable and reconfigurable optical hardware for deep learning," Sci. Rep. **11**(1), 3144 (2021).
31. C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, and H. Chen, "Performing optical logic operations by a diffractive neural network," Light: Sci. Appl. **9**(1), 59 (2020).
32. S. L. Suárez Gómez, C. González-Gutiérrez, E. Díez Alonso, J. D. Santos Rodríguez, M. L. Sánchez Rodríguez, J. Carballido Landeira, A. Basden, and J. Osborn, "Improving adaptive optics reconstructions with a deep learning approach," in *Hybrid Artificial Intelligent Systems*, F. J. de Cos Juez, J. R. Villar, E. A. de la Cal, Á. Herrero, H. Quintián, J. A. Sáez, and E. Corchado, eds. (Springer International Publishing, Cham, 2018), pp. 74–83.
33. K. Yanny, K. Monakhova, R. W. Shuai, and L. Waller, "Deep learning for fast spatially varying deconvolution," Optica **9**(1), 96–99 (2022).
34. G. Wetzstein, H. Ikoma, C. Metzler, and Y. Peng, "Deep optics: Learning cameras and optical computing systems," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, (2020), pp. 1313–1315.
35. Y. E. Peng, A. Veeraraghavan, W. Heidrich, and G. Wetzstein, "Deep optics: Joint design of optics and image recovery algorithms for domain specific cameras," in *ACM SIGGRAPH 2020 Courses*, (Association for Computing Machinery New York, NY, USA, 2020), SIGGRAPH '20.
36. J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019).
37. C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020).
38. Y. Liu, C. Zhang, T. Kou, Y. Li, and J. Shen, "End-to-end computational optics with a singlet lens for large depth-of-field imaging," Opt. Express **29**(18), 28530–28548 (2021).
39. Q. Sun, C. Wang, Q. Fu, X. Dun, and W. Heidrich, "End-to-end complex lens design with differentiate ray tracing," ACM Trans. Graph. **40**(4), 1–13 (2021).
40. E. Nehme, B. Ferdman, L. E. Weiss, T. Naor, D. Freedman, T. Michaeli, and Y. Shechtman, "Learning optimal wavefront shaping for multi-channel imaging," IEEE Trans. Pattern Anal. Mach. Intell. **43**(7), 2179–2192 (2021).
41. E. Tseng, S. Colburn, J. Whitehead, L. Huang, S.-H. Baek, A. Majumdar, and F. Heide, "Neural nano-optics for high-quality thin lens imaging," Nat. Commun. **12**(1), 6493 (2021).
42. N. Yu and F. Capasso, "Flat optics with designer metasurfaces," Nat. Mater. **13**(2), 139–150 (2014).
43. A. Pors, O. Albrektsen, I. P. Radko, and S. I. Bozhevolnyi, "Gap plasmon-based metasurfaces for total control of reflected light," Sci. Rep. **3**(1), 2155 (2013).
44. J. W. Goodman, *Introduction to Fourier optics* (Roberts and Company Publishers, 2005), 3rd ed.
45. A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. rep., Department of Computer Science, University of Toronto (2009). https://www.cs.toronto.edu/kriz/learning-features-2009-TR.pdf.