HHS Public Access

Author manuscript

*Nat Nanotechnol*. Author manuscript; available in PMC 2024 April 20.

Published in final edited form as:

*Nat Nanotechnol*. 2024 April ; 19(4): 471–478. doi:10.1038/s41565-023-01557-2.

# Multichannel meta-imagers for accelerating machine vision

**Hanyu Zheng**[1], **Quan Liu**[2], **Ivan I. Kravchenko**[3], **Xiaomeng Zhang**[4], **Yuankai Huo**[2], **Jason G. Valentine**[4,✉]

[1]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN, USA.

[2]Department of Computer Science, Vanderbilt University, Nashville, TN, USA.

[3]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA.

[4]Department of Mechanical Engineering, Vanderbilt University, Nashville, TN, USA.

## Abstract

Rapid developments in machine vision technology have impacted a variety of applications, such as medical devices and autonomous driving systems. These achievements, however, typically necessitate digital neural networks with the downside of heavy computational requirements and consequent high energy consumption. As a result, real-time decision-making is hindered when computational resources are not readily accessible. Here we report a meta-imager designed to work together with a digital back end to offload computationally expensive convolution operations into high-speed, low-power optics. In this architecture, metasurfaces enable both angle and polarization multiplexing to create multiple information channels that perform positively and negatively valued convolution operations in a single shot. We use our meta-imager for object classification, achieving 98.6% accuracy in handwritten digits and 88.8% accuracy in fashion images. Owing to its compactness, high speed and low power consumption, our approach could find a wide range of applications in artificial intelligence and machine vision applications.

The rapid development of digital neural networks and the availability of large training datasets have enabled a wide range of machine-learning-based applications, including image analysis[1,2], speech recognition[3,4] and machine vision[5]. However, enhanced performance is typically associated with a rise in model complexity, leading to larger computing



✉**Correspondence and requests for materials** should be addressed to Jason G. Valentine. jason.g.valentine@vanderbilt.edu.

Author contributions

H.Z. and J.G.V. developed the idea. H.Z. conducted the optical modelling and system design. Q.L. and H.Z. trained the digital neural network. H.Z. fabricated the samples. I.I.K. performed the silicon growth and electron-beam-lithography for the metasurfaces. H.Z. conducted the experimental measurements. H.Z., Q.L. and X.Z. performed the data analysis. H.Z. and J.G.V. wrote the manuscript with input from all the authors. The project was supervised by Y.H. and J.G.V.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-023-01557-2.

Competing interests

The authors declare no competing interests.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41565-023-01557-2.

requirements[6]. The escalating use and complexity of neural networks have resulted in increases in energy consumption and limiting real-time decision-making when large computational resources are not readily accessible. These issues are especially critical to the performance of machine vision[7–9] in autonomous systems where the imager and processor must have a small size, weight and power consumption for onboard processing and still maintain low latency, high accuracy and highly robust operation. These opposing requirements necessitate the development of new hardware and software solutions as the demands on machine vision systems continue to grow.

Optics has long been studied as a way to speed up computational operations as well as increase energy efficiency[10–16]. In accelerating vision systems, there is the unique opportunity to offload computation into the front-end imaging optics by designing an imager that is optimized for a particular computational task. A free-space optical computational, based on Fourier optics[17–20], actually predates modern digital circuitry and allows for the highly parallel execution of convolution operations, which comprise a majority of floating-point operations in machine vision architectures[21,22]. The challenge with Fourier-based processors is that they are traditionally employed by reprojecting the imagery using spatial light modulators and coherent sources, enlarging the system size compared with chip-based approaches[23–28]. Although coherent illumination is not strictly required, it allows for more freedom in convolution operations including the ability to achieve the negatively valued kernels needed for spatial derivatives. Optical diffractive neural networks[29–31] offer an alternative approach even though they are also employed with coherent sources and thus are best suited as back-end processors with image data being reprojected.

Metasurfaces offer a unique platform for implementing front-end optical computation as they can reduce the size of the optical elements and allow for a wider range of optical properties including polarization[32,33], wavelength[34,35] and angle of incidence[36,37] to be utilized in computation. For instance, metasurfaces have been demonstrated with angle-of-incidence-dependent transfer functions for realizing compact optical differentiation systems[38–41] with no need to pass through the Fourier plane of a two-lens system. In addition, wavelength-multiplexed metasurfaces, combined with optoelectronic subtraction, have been used to achieve negatively valued kernels for executing single-shot differentiation with incoherent light[42,43]. Differentiation, however, is a single convolution operation whereas most machine vision systems require multiple independent channels. There have been recent studies on multichannel convolutional front ends, but these have been limited in transmission efficiency and computational complexity, achieving only positively valued kernels with a stride that is equal to the kernel size, preventing the implementation of common digital designs[44,45]. Although these are important steps towards a computational front end, an architecture is still needed for generating the multiple independent, and arbitrary, convolution channels that are used in machine vision systems.

Here we demonstrate a meta-imager that can serve as a multichannel convolutional accelerator for incoherent light. To achieve this, the point spread function (PSF) of the imaging meta-optic is engineered to achieve parallel multichannel convolution using a single aperture implemented with angular multiplexing (Fig. 1). In addition, positively and negatively valued kernels are achieved for incoherent illumination by

using polarization multiplexing[46], combined with a polarization-sensitive camera and optoelectronic subtraction. A second metasurface corrector is also employed to widen the field of view (FOV) for imaging objects in the natural world and both metasurfaces are restricted to phase functions, yielding high transmission efficiency. As a proof of concept, the platform is used to experimentally demonstrate the classification of modified National Institute of Standards and Technology (MNIST) and Fashion MNIST datasets[47] with measured accuracies of 98.6% and 88.8%, respectively. In both cases, 94% of the operations are offloaded from the digital platform into the front-end optics.

## Angular and polarization multiplexing

The meta-optic described here is designed to optically implement the convolutional layers at the front end of a digital neural network. In a digital network, convolution comprises matrix multiplication of the object image and an $N \times N$ pixel kernel, with each pixel having an independent weight, as illustrated for $N = 3$ (Fig. 2a). The kernel is multiplied over an area of the image using a dot product and then rastered across the image, moving by a single pixel at each step until it is swept across the entire image, forming a single feature map. Under incoherent illumination, the optical convolutional is expressed as Image = Object $\otimes$ |PSF($x$, $y$)|, where PSF($x$, $y$) is the PSF of the optics. Typically, in implementing the optical version of digital convolution, PSF($x$, $y$) is a continuous function that is discretized in forming the digital kernel. Here we take a different approach, creating a true optical analogue to the digital kernel. This is done by engineering the PSF($x$, $y$) value (Fig. 2a) to possess $N \times N$ focal spots, each with a different weight, or image intensity, that matches the desired digital kernel weight. These focal spots will result in $N \times N$ images of the object being formed, which are spatially overlapped on the sensor and offset based on the separation in the focal-spot positions. In this case, we are rastering the weighted images with the summing operation in the dot product being achieved by overlapping the images on the camera.

In this architecture, positively and negatively valued kernel weights are achieved by encoding the focal spots with either right-hand circular polarization (RCP) or left-hand circular polarization (LCP), respectively. The circularly polarized signal is decoded by using a quarter-wave plate combined with a polarization-sensitive camera containing four directional gratings integrated onto each pixel. The RCP- and LCP-encoded feature maps (Fig. 2a) are then independently recorded using the polarization-sensitive camera, with summing being achieved by digitally subtracting the LCP feature map from the RCP feature map. The convolution generated by this method is identical to the digital process, which is evidenced by comparing the digital and optical feature maps (Fig. 2a). We have used this approach for several reasons. First, as explained later, the phase and amplitude profiles associated with our desired PSF($x$, $y$) is analytical, substantially simplifying the design process and allowing us to achieve numerous independent feature maps, or channels, using one aperture. In addition, since we have a true optical analogue to a digital system, we can directly implement digital kernel designs with optics, removing the optic stage from the design loop, further speeding up the design process. To achieve the desired optical response, we employ a bilayer-metasurface architecture (Fig. 2b). In this architecture, the first metasurface splits the incident signal into angular channels of varying weights, whereas

birefringence in this layer is used to encode the positive and negative kernel values in the RCP and LCP cases, respectively. The second metasurface is polarization insensitive and serves as the focusing optic to create an $N \times N$ focal-spot array for each channel.

## Meta-optic design

Meta-optic design began by optimizing a two-metasurface lens, comprising a wavefront corrector and a focuser, to be coma free over a ±10° angular range using the commercial software Zemax (Methods). Supplementary Note 1 provides the phase profiles and angular response of the metasurfaces, which shows a constant focal-spot shape within the designed angular range. A wider FOV can be achieved by further cascading metasurfaces (Supplementary Note 2). Once the coma-free meta-optic was designed, angular multiplexing was applied to the first metasurface to form focal-spot arrays as the convolution kernels. The focal-spot position is controlled using angular multiplexing, with each angle corresponding to a kernel pixel. By encoding a weight to each angular component, the system PSF, serving as the optical kernel, can be readily engineered. The analytical expression of the complex-amplitude profile multiplexing all the angular signals is given by

$$A(x,y)=\sum_{m}^{M}\sum_{n}^{N}\sqrt{w_{mn}}\exp\left\{\mathrm{i}\frac{2\pi}{\lambda}\left[x\sin\left(\theta_{x\,|\,mn}\right)+y\sin\left(\theta_{y\,|\,mn}\right)\right]\right\} \tag{1}$$

where $A(x, y)$ is a complex-amplitude field. Also, $M$ and $N$ denote the row and column number of the elements in the kernel, respectively; $w_{mn}$ is the corresponding weight of each element, which is normalized to a range of [0, 1]; $\lambda$ is the working wavelength; $x$ and $y$ are spatial coordinates; and $\theta_{x/mn}$ and $\theta_{y/mn}$ are the designed angles with a small variation to form the kernel elements. The deflection angles are selected to realize the desired PSF for incoherent light illumination, which is given by

$$PSF(x,y)=\sum_{m}^{M}\sum_{n}^{N}w_{mn}\Theta\left\{x-f_1c\left[\frac{x_0}{f_2}+\tan\left(\theta_{x\,|\,mn}\right)\right]\right.$$
$$\left.y-f_1c\left[\frac{y_0}{f_2}+\tan\left(\theta_{y\,|\,mn}\right)\right]\right\} \tag{2}$$

where $x_0$ and $y_0$ are the location of the object and $\Theta(x, y)$ is the focal spot excited by a plane wave. Also, $f_1$ is the focal length of the meta-imager, whereas $c$ is a constant fitted based on the imaging system; $f_2$ is the distance from the object to the front aperture. Supplementary Note 3 provides the detailed derivation. The separation distance of each focal spot, $\Delta p$, defines the imaged pixel size of the object. Based on the prescribed PSF, the required angles $\theta_{mn}$ can be derived from equation (2), which can be further extended into an off-axis imaging case (Supplementary Note 4) for the purpose of multichannel, single-shot convolutional applications.

In equation (1), we employ a spatially varying complex-valued amplitude function (Supplementary Note 5 shows the workflow of the design process) that would ultimately introduce a large reflection loss, leading to a low diffraction efficiency[48]. To overcome this limitation, an optimization platform was developed based on the angular spectrum propagation method and stochastic gradient descent solver, which converts the complex-amplitude profile into a phase-only metasurface. The algorithm encodes a phase term, $\exp(i\phi_{mn})$, onto each weight $w_{mn}$ based on the loss function, namely, $\mathcal{F} = \sum\left(|A|^2 - I\right)^2/N$. Here $I$ is a matrix consisting of unity elements and $N$ is the total pixel number. The intensity profile becomes more consistent and closer to a phase-only device by minimizing the loss function during optimization (Supplementary Note 6 shows the detailed algorithm). The phase-only approximation can effectively avoid loss in the complex-amplitude function, leading to a theoretical diffraction efficiency as high as 84.3% where 14.0% of the loss is introduced by Fresnel reflection, which can be removed by adding antireflection coatings.

## Hybrid neural network for object classification

To validate the performance of this architecture, a shallow convolutional neural network was trained for the purpose of image classification. The neural network architecture (Fig. 3a) contains an optical convolution layer followed by digital max pooling, a rectified linear unit activation function and a fully connected layer. In the convolution process, 12 independent kernels are used to extract the feature maps and the overall intensity of positive and negative channels was set to be equal due to energy conservation from the phase-only approximation in the meta-optic design. Since neural network training is a high-dimensional problem with infinite solutions, the above kernel restrictions do not notably affect the final performance (Supplementary Note 7). Each kernel comprised $N = 7$ pixels instead of a more typical $N = 3$ format, to correlate neurons within a broader FOV[49], leading to better performance for large-scale object recognition. Methods describes the detailed training process. To finish the classification, the feature maps extracted by the compound meta-optic are fed into the digital component of the neural network. In this architecture, 94% of the total operations are offloaded from the digital platform into the meta-optic, leading to a substantial speed up for classification tasks (Supplementary Note 8).

## Meta-optic implementation

To realize the first polarization-selective metasurface, elliptical nanopillars were chosen as the base meta-atoms (Fig. 3b). The width and length of the nanopillars were designed so that the nanopillars serve as half-wave plates. This choice introduces a spin-decoupled phase response by simultaneously introducing geometrical and a locally resonant phase delay; hence, independent phase control over orthogonal circularly polarized states can be achieved. The analytical expression of the phase delay for the different polarization states is described as

$$\begin{bmatrix}\phi_{\mathrm{LCP}}\\ \phi_{\mathrm{RCP}}\end{bmatrix} = \begin{bmatrix}\phi_x + 2\theta + \pi/4\\ \phi_x - 2\theta - \pi/4\end{bmatrix} \quad (3)$$

Here $\phi_x$ is the phase delay of the meta-atoms along the $x$ axis at $\theta = 0$. Hence, by tuning the length, width and rotation angle, the phase delay of LCP and RCP light can be independently controlled (Supplementary Note 9 shows the detailed derivation). The second metasurface was designed based on circular nanopillars arranged in a hexagonal lattice for realizing polarization-insensitive phase control. Supplementary Note 10 shows the phase delay of the circular nanopillars as a function of diameter.

## Fabrication and characterization of meta-optic

Two versions of the meta-optic classifier were fabricated based on networks trained for the MNIST and Fashion MNIST datasets (Supplementary Note 11 shows one set of the phase profiles). The fabrication of the meta-optic began with a silicon device layer on a fused silica substrate patterned by standard electron-beam lithography followed by reactive ion etching. A thin polymethyl methacrylate layer was spin coated over the device as the protective and index-matching layer. Methods describes the detailed fabrication process. An optical image of the two metasurfaces comprising the meta-optic is exhibited in Fig. 4a,b, with the inset showing the meta-atoms. To align the compound meta-optic, the first metasurface was mounted in a rotational stage (CRM1PT, Thorlabs), whereas the second layer was fitted in a three-axis translational stage (CXYZ05A, Thorlabs). The metasurfaces are aligned in situ and characterized in a cage system (Supplementary Note 12 shows the detailed alignment setup). A meta-hologram was fabricated on the first layer alongside the device to assist the alignment process by forming an alignment pattern at a prescribed distance along the optical axis corresponding to the designed separation distance. The alignment process was finished by overlapping the alignment pattern with the low-transmission register on the second layer. Due to the large size (millimetre scale) of each metasurface layer, the meta-optic exhibits a high alignment tolerance. The system performance remains constant under a horizontal misalignment of 65 μm and vertical displacement of ±400 μm, indicating the robustness of the entire convolutional system. Supplementary Note 13 shows the alignment error analysis.

To characterize the optical properties of the fabricated meta-optic, a linearly polarized laser was used for illumination in obtaining the PSF (Supplementary Note 14 shows the detailed characterization setup). The linearly polarized light source includes the LCP and RCP components with equal strength. The PSF at the focal plane of the compound meta-optic (Fig. 4c,d) indicates a good match between the ideal and measured results, where the red and blue colours represent positive and negative values, respectively.

Optical convolution of a grey-scale Vanderbilt logo was used to characterize the accuracy of the fabricated meta-optic (Fig. 4e). To accomplish this, an imaging system using a liquid-crystal-based spatial light modulator was built (Supplementary Note 15). An incoherent tungsten lamp with a 10-nm-wide bandpass filter was used for spatial light modulator illumination. The feature maps extracted by the meta-optic were recorded by a polarization-sensitive camera (DZK 33UX250, Imaging Source) where orthogonally polarized channels are simultaneously recorded using polarization filters on each camera pixel. A comparison between the digital and measured feature maps, recorded on the camera, is illustrated in Fig. 4e. The pixel intensity from the digital and measured convolutional results at the same position were extracted and compared to evaluate the convolution fidelity. The deviation

between the ideal and measured results, defined by $\sigma = \sum_{n=1}^{N} |D_{i,n} - D_{m,n}|/(2N)$, was calculated as 3.83%, where $D_i$ and $D_m$ are the ideal and measured intensity, respectively, and $N$ is the number of total pixels. The error originates from stray light, fabrication imperfections, local phase approximation and metasurface misalignment (Supplementary Note 16 shows the detailed system error analysis). These errors also result in a small amount of zeroth-order diffracted light being introduced from the first metasurface, leading to a spot at the centre of the imaging plane. However, the polarization state of the zeroth-order light remains unchanged, with the energy evenly distributed in the two circularly polarized channels. Hence, subtraction between the information channels allows the zeroth-order pattern to be cancelled, not affecting the classification performance. Supplementary Note 17 provides a detailed discussion.

## Object classification for machine vision

As a proof of concept in demonstrating multichannel convolution, a full meta-optic classifier was first designed and fabricated based on the classification of the MNIST dataset, which includes 60,000 handwritten digit training images with a 28 × 28 pixel format. The feature maps of 1,000 digits, not in the training set, were extracted using the meta-optic to characterize the system performance. An example input image is exhibited in Fig. 5a, with the corresponding feature maps shown in Fig. 5b. Supplementary Note 18 shows the kernels and feature maps for all the channels. The measured feature maps match well with the theoretical prediction (Fig. 5b), indicating good fidelity in the optical convolution process. The theoretical and experimental confusion matrices for this testing dataset are shown in Fig. 5c, demonstrating 99.3% accurate classification in theory and 98.6% accurate classification in the measurement. The small drop in accuracy probably results from the small inaccuracy in the realized optical kernels. Although the system was designed at a single wavelength, simulations indicate a minimal accuracy drop up to an illumination bandwidth of 50 nm, indicating that the experimental bandwidth of 10 nm should have a minimal impact (Supplementary Note 19).

To explore the flexibility of the approach, a dataset with higher spatial frequency information, namely, the Fashion MNIST dataset, was used for training the model with an example input image provided in Fig. 5d. This dataset includes 60,000 training images of clothing articles that contain images with higher spatial frequencies than the MNIST handwritten digit dataset. The ideal and measured feature maps are compared in Fig. 5e, indicating good agreement. Supplementary Note 20 shows all of the designed kernel profiles and feature maps. The confusion matrices for Fashion MNIST are illustrated in Fig. 5f, with 90.2% accurate classification in theory and 88.8% in measurement. To validate the role of the optical convolution layer, a reference model for the MNIST handwritten digit classification, without a convolutional layer, was trained, resulting in an accuracy of 80.3%, illustrating the importance of the convolution operations (Supplementary Note 21). Compared with the MNIST dataset, the Fashion MNIST model has a slightly lower accuracy, in theory, due to the higher resolution features in the dataset. Specifically, for class 7 in the Fashion MNIST dataset, the accuracy predicted by the optical front end dropped from 81.4% to 67.0%, with the model misidentifying the images as classes 1, 3, 4 and 5. We expect these classes to share the same features during model training (Supplementary

Note 22). These mixed features can be potentially distinguished by adaptively tuning the loss function during model training[50] or utilizing novel neural network architecture such as a vision transformer[51] with better performance at comparable floating-point operations.

To understand the scalability of the meta-imager, the accuracy of classification as a function of the areal density of the basic computing unit was calculated (Fig. 5g). The optical computing unit density is defined as the convolutional pixels per unit area, where we assume each convolutional pixel is matched to a physical pixel on a photodetector. The pixel size is dictated by the separation distance between the neighbouring focal spots in the PSF, which is ultimately dictated by the diffraction limit. The prediction accuracy is based on the MNIST dataset and the theoretical accuracy remains as high as ~99% until the pixel size drops below 2 μm, at which point the neighbouring focal spots are below the diffraction limit, resulting in additional aberration in the output features (Fig. 5g, inset). Thus, although a pixel size of 12 μm is demonstrated in this work as a proof of concept, the system functionality would remain unchanged, in theory, with up to six times higher areal computing unit density. For perspective, the meta-imager computing unit density can be compared with the multiply–accumulation unit density and size based on the current 7-nm-node architecture[52], which results in multiply–accumulation units with a size of ~7 μm × 7 μm.

## Conclusions

Our meta-imager is a proof of concept for a convolutional front end that can be used to replace the traditional imaging optics in machine vision applications, encoding information in a more efficient basis for back-end processing. In this context, negatively valued kernels and multichannel convolution, enabled by the meta-optic, allows one to increase the number of operations that can be offloaded into the front-end optics. Furthermore, the architecture allows for incoherent illumination and a reasonably wide FOV, both of which are needed for implementation in imaging natural scenes with ambient illumination. Although a tradeoff exists between the channel number and the viewing-angle range, a multiaperture architecture could be designed without deteriorating the FOV in a single imaging channel[53]. In addition, we have not attempted to optimize the operation bandwidth, which could be addressed through dispersion engineering, over modest apertures, combination with broadband refractive optics or use of dispersion to perform wavelength-dependent functions. Further acceleration can be realized via the integration of a meta-imager front end directly with a chip-based photonics back end such that data readout and transport can be achieved without analogue-to-digital converters for ultrafast and low-latency processing.

Our meta-imager does put restrictions on the depth, or number of layers, in the optical front end, which means that it provides the most benefit in lightweight neural networks such as those found in power-limited or high-speed autonomous applications. Recent advances in machine learning, such as the use of larger kernels for network layer compression[54] and reparameterization[55], could further improve the effectiveness of single-layer, or few-layer, meta-imager front ends. In addition, the capability of the meta-optic for multifunctional processing, including wavelength- and polarization-based discrimination, can be used to further increase information collection[44]. As a result, this general architecture for meta-imagers can be highly parallel and bridge the gap between the natural world and

digital systems, potentially finding use beyond machine vision[56] in applications such as information security[57,58] and quantum communications[59].

## Methods

### Optimization of coma-free meta-optic

The coma-free meta-optic contains two metasurfaces, whose phase profiles were optimized by the ray-tracing technique using a commercial optical design software (Zemax OpticStudio). The phase profile of each layer was defined by even-order polynomials according to the radial coordinate $\rho$ as follows:

$$\phi(\rho) = \sum_{n=1}^{5} a_n \left(\frac{\rho}{R}\right)^{2n} \quad (4)$$

where $R$ is the radius of the metasurface and $a_n$ is the optimized coefficient to minimize the focal-spot size of the bilayer metasurfaces system under an incidence angle of up to 13°. The diameter of the second layer of the metasurface was 1.5 times that of the first layer to capture all the light under high-incidence-angle illumination. The phase profiles were then wrapped within 0 to 2π to be fitted by meta-atoms.

### Digital neural network training

The MNIST and Fashion MNIST databases, each containing 60,000 training images with the 28 × 28 pixel format, were used to train the digital convolutional neural network. The channel number for convolution was set to 12, whereas the kernel size was fixed at 7 × 7, with the size of the convolutional result remaining the same. The details of the neural network architecture are shown in Fig. 3a. During forward propagation in the neural network, an additional loss function defined by $\mathscr{L} = \sum_{n=1}^{N} w_n$ was added to ensure an equal total intensity of positive and negative kernel values, where $w_n$ is the weight of each kernel. All the kernel values are normalized to [−1, 1], by dividing with a constant, to maximize the diffraction efficiency in the optics. An Adam optimizer was utilized for training the digital parameters with a learning rate of 0.001. The training process is sustained over 50 epochs, during which the performance is optimized by minimizing the negative log-likelihood loss from comparing prediction probabilities and ground-truth labels. The algorithm was programmed based on PyTorch 1.10.1 and CUDA 11.0 with a Quadro RTX 5000/PCIe/SSE2 as the graphics card.

### Numerical simulation

The complex transmission coefficients of the silicon nanopillars were calculated using an open-source rigorous coupled-wave analysis solver called RETICOLO[60]. A square lattice with a period of 0.45 μm was used for the first metasurface with a working wavelength at 0.87 μm. The second metasurface was assigned a hexagonal lattice with a period of 0.47 μm. During full-wave simulation, the indices of silicon and fused silica characterized by ellipsometry were set at 3.74 and 1.45, respectively.

### Metasurface fabrication

Electron-beam-lithography-based lithography was used to fabricate all the metasurface layers. First, low-pressure chemical vapour deposition was utilized to deposit a 630-nm-thick silicon device layer on a fused silica substrate. A polymethyl methacrylate photoresist was then spin coated on the silicon layer, followed by thermal evaporation of a 10-nm-thick Cr conduction layer. The electron-beam lithography system then exposed the photoresist, and after removing the Cr layer, the pattern was developed by a methyl isobutyl ketone/iso-propyl alcohol solution. A 30 nm $Al_2O_3$ hard mask was deposited via electron-beam evaporation, followed by a lift-off process with an *N*-methyl-2-pyrrolidone solution. The silicon was then patterned using reactive ion etching, and a 1-μm-thick layer of polymethyl methacrylate was spin coated to encase the nanopillar structures as a protective and index-matching layer.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

H.Z. and J.G.V. acknowledge support from DARPA under contract HR001118C0015 and NAVAIR under contract N6893622C0030. X.Z. acknowledges support from ONR under contract N000142112468. Y.H. and Q.L. acknowledge support from NIH under contract R01DK135597. Meta-optic devices were manufactured as part of a user project at the Center for Nanophase Materials Sciences (CNMS), which is a US Department of Energy, Office of Science User Facility, Oak Ridge National Laboratory.

## Data availability

The data that support the findings of this study are available in the Article and its Supplementary Information and/or are available from the corresponding author upon reasonable request.

## References

1. Simonyan K & Zisserman A Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations 1–14 (ICLR, 2015).
2. Wang G et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE Trans. Med. Imaging 37, 1562–1573 (2018). [PubMed: 29969407]
3. Furui S, Deng L, Gales M, Ney H & Tokuda K Fundamental technologies in modern speech recognition. IEEE Signal Process Mag. 29, 16–17 (2012).
4. Sak H, Senior A, Rao K & Beaufays F Fast and accurate recurrent neural network acoustic models for speech recognition. In Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH 1468–1472 (ISCA, 2015).
5. He K, Zhang X, Ren S & Sun J Deep residual learning for image recognition. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 770–778 (IEEE, 2016).
6. Lecun Y, Bengio Y & Hinton G Deep learning. Nature 521, 436–444 (2015). [PubMed: 26017442]
7. Mennel L et al. Ultrafast machine vision with 2D material neural network image sensors. Nature 579, 62–66 (2020). [PubMed: 32132692]
8. Liu L et al. Computing systems for autonomous driving: state of the art and challenges. IEEE Internet Things J. 8, 6469–6486 (2021).
9. Shi W et al. LOEN: lensless opto-electronic neural network empowered machine vision. Light Sci. Appl 11, 121 (2022). [PubMed: 35508469]

10. Hamerly R, Bernstein L, Sludds A, Soljačić M & Englund D Large-scale optical neural networks based on photoelectric multiplication. Phys. Rev. X 9, 021032 (2019).
11. Wetzstein G et al. Inference in artificial intelligence with deep optics and photonics. Nature 588, 39–47 (2020). [PubMed: 33268862]
12. Shastri BJ et al. Photonics for artificial intelligence and neuromorphic computing. Nat. Photon 15, 102–114 (2021).
13. Xue W & Miller OD High-NA optical edge detection via optimized multilayer films. J. Optics 23, 125004 (2021).
14. Wang T et al. An optical neural network using less than 1 photon per multiplication. Nat. Commun 13, 123 (2022). [PubMed: 35013286]
15. Wang T et al. Image sensing with multilayer nonlinear optical neural networks. Nat. Photon 17, 8–17 (2023).
16. Badloe T, Lee S & Rho J Computation at the speed of light: metamaterials for all-optical calculations and neural networks. Adv. Photon 4, 064002 (2022).
17. Vanderlugt A Optical Signal Processing (Wiley, 1993).
18. Chang J, Sitzmann V, Dun X, Heidrich W & Wetzstein G Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. Sci. Rep 8, 12324 (2018). [PubMed: 30120316]
19. Colburn S, Chu Y, Shilzerman E & Majumdar A Optical frontend for a convolutional neural network. Appl. Opt 58, 3179 (2019). [PubMed: 31044792]
20. Zhou T et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit. Nat. Photon 15, 367–373 (2021).
21. Chen YH, Krishna T, Emer JS & Sze V Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J. Solid-State Circuits 52, 127–138 (2017).
22. Neshatpour K, Homayoun H & Sasan A ICNN: the iterative convolutional neural network. In ACM Transactions on Embedded Computing Systems 18, 119 (ACM, 2019).
23. Xu X et al. 11 TOPS photonic convolutional accelerator for optical neural networks. Nature 589, 44–51 (2021). [PubMed: 33408378]
24. Feldmann J et al. Parallel convolutional processing using an integrated photonic tensor core. Nature 589, 52–58 (2021). [PubMed: 33408373]
25. Wu C et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. Nat. Commun 12, 96 (2021). [PubMed: 33398011]
26. Zhang H et al. An optical neural chip for implementing complex-valued neural network. Nat. Commun 12, 457 (2021). [PubMed: 33469031]
27. Ashtiani F, Geers AJ & Aflatouni F An on-chip photonic deep neural network for image classification. Nature 606, 501–506 (2022). [PubMed: 35650432]
28. Fu T et al. Photonic machine learning with on-chip diffractive optics. Nat. Commun 14, 70 (2023). [PubMed: 36604423]
29. Lin X et al. All-optical machine learning using diffractive deep neural networks. Science 361, 1004–1008 (2018). [PubMed: 30049787]
30. Qian C et al. Performing optical logic operations by a diffractive neural network. Light Sci. Appl 9, 59 (2020). [PubMed: 32337023]
31. Luo X et al. Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. Light Sci. Appl 11, 158 (2022). [PubMed: 35624107]
32. Kwon H, Arbabi E, Kamali SM, Faraji-Dana MS & Faraon A Single-shot quantitative phase gradient microscopy using a system of multifunctional metasurfaces. Nat. Photon 14, 109–114 (2020).
33. Xiong B et al. Breaking the limitation of polarization multiplexing in optical metasurfaces with engineered noise. Science 379, 294–299 (2023). [PubMed: 36656947]
34. Khorasaninejad M et al. Metalenses at visible wavelengths: diffraction-limited focusing and subwavelength resolution imaging. Science 352, 1190–1194 (2016). [PubMed: 27257251]
35. Kim J et al. Scalable manufacturing of high-index atomic layer–polymer hybrid metasurfaces for metaphotonics in the visible. Nat. Mater 22, 474–481 (2023). [PubMed: 36959502]

36. Levanon N et al. Angular transmission response of in-plane symmetry-breaking quasi-BIC all-dielectric metasurfaces. ACS Photonics 9, 3642–3648 (2022).
37. Nolen JR, Overvig AC, Cotrufo M & Alù A Arbitrarily polarized and unidirectional emission from thermal metasurfaces. Preprint at https://arxiv.org/abs/2301.12301 (2023).
38. Guo C, Xiao M, Minkov M, Shi Y & Fan S Photonic crystal slab Laplace operator for image differentiation. Optica 5, 251–256 (2018).
39. Cordaro A et al. High-index dielectric metasurfaces performing mathematical operations. Nano Lett. 19, 8418–8423 (2019). [PubMed: 31675241]
40. Zhou Y, Zheng H, Kravchenko II & Valentine J Flat optics for image differentiation. Nat. Photon 14, 316–323 (2020).
41. Fu W et al. Ultracompact meta-imagers for arbitrary all-optical convolution. Light Sci. Appl 11, 62 (2022). [PubMed: 35304870]
42. Wang H, Guo C, Zhao Z & Fan S Compact incoherent image differentiation with nanophotonic structures. ACS Photonics 7, 338–343 (2020).
43. Zhang X, Bai B, Sun HB, Jin G & Valentine J Incoherent optoelectronic differentiation based on optimized multilayer films. Laser Photon Rev. 16, 2200038 (2022).
44. Zheng H et al. Meta-optic accelerators for object classifiers. Sci. Adv 8, eabo6410 (2022). [PubMed: 35895828]
45. Bernstein L et al. Single-shot optical neural network. Sci. Adv 9, eadg7904 (2023). [PubMed: 37343096]
46. Shen Z et al. Monocular metasurface camera for passive single-shot 4D imaging. Nat. Commun 14, 1035 (2023). [PubMed: 36823191]
47. LeCun Y, Bottou L, Bengio Y & Haffner P Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2323 (1998).
48. Zheng H et al. Compound meta-optics for complete and loss-less field control. ACS Nano 16, 15100–15107 (2022). [PubMed: 36018810]
49. Liu S et al. More ConvNets in the 2020s: scaling up kernels beyond 51×51 using sparsity. In 11th International Conference on Learning Representations 1–23 (ICLR, 2023).
50. Barron JT A general and adaptive robust loss function. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 4326–4334 (IEEE, 2019).
51. Dosovitskiy A et al. An image is worth 16×16 words: transformers for image recognition at scale. In 9th International Conference on Learning Representations 1–22 (ICLR, 2021).
52. Stillmaker A & Baas B Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm. Integration 58, 74–81 (2017).
53. McClung A, Samudrala S, Torfeh M, Mansouree M & Arbabi A Snapshot spectral imaging with parallel metasystems. Sci. Adv 6, eabc7646 (2020). [PubMed: 32948595]
54. Ding X, Zhang X, Han J & Ding G Scaling up your kernels to 31 × 31: revisiting large kernel design in CNNs. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 11953–11965 (IEEE, 2022).
55. Ding X et al. RepVgg: making VGG-style ConvNets great again. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 13728–13737 (IEEE, 2021).
56. Li L et al. Intelligent metasurface imager and recognizer. Light Sci. Appl 8, 97 (2019). [PubMed: 31645938]
57. Zhao R et al. Multichannel vectorial holographic display and encryption. Light Sci. Appl 7, 95 (2018). [PubMed: 30510691]
58. Kim I et al. Pixelated bifunctional metasurface-driven dynamic vectorial holographic color prints for photonic security platform. Nat. Commun 12, 3614 (2021). [PubMed: 34127669]
59. Li L et al. Metalens-array-based high-dimensional and multiphoton quantum source. Science 368, 1487–1490 (2020). [PubMed: 32587020]
60. Hugonin AJP & Lalanne P RETICOLO software for grating analysis. Preprint at https://arxiv.org/abs/2101.00901 (2023).

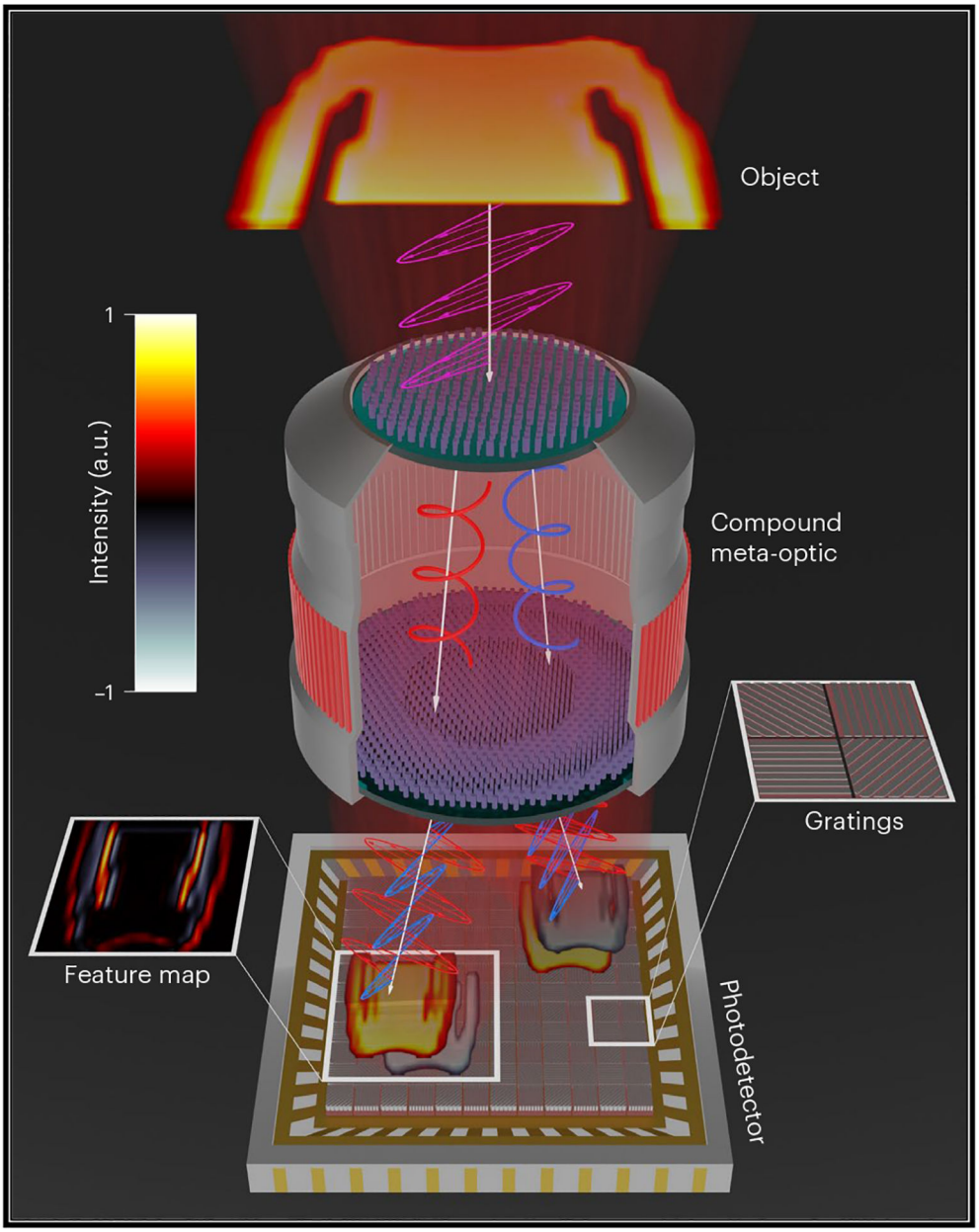


**Fig. 1 |. Schematic of the meta-imager.**

The meta-imager enables multichannel signal processing for replacing convolution operations in a digital neural network. A bilayer meta-optic system encoded by the predesigned kernels is utilized to achieve optical convolution with the incoherent light source to be used for object illumination. The positive and negative values are distinguished and recorded as feature maps by a polarization-sensitive photodetector, where an oriented grating sits on each photodetector pixel for polarized signal sorting.

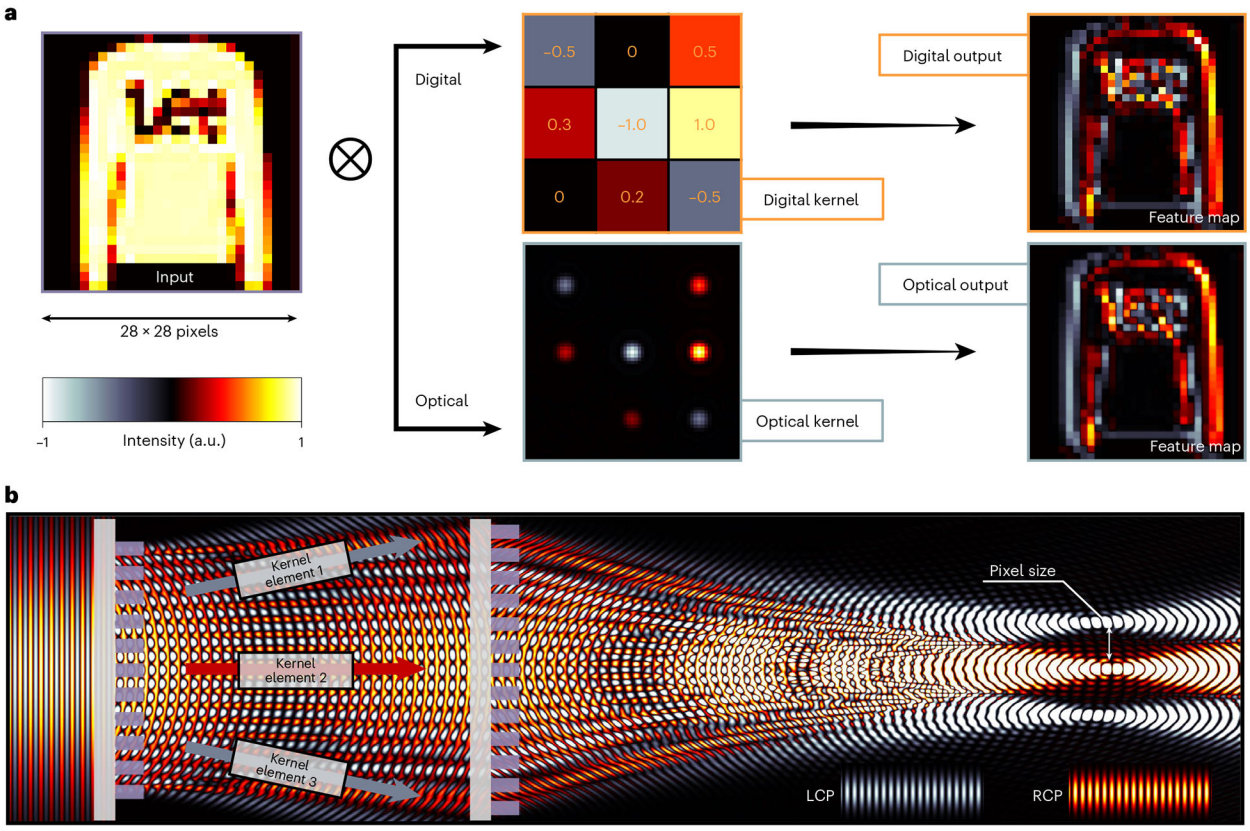


**Fig. 2 |. Meta-optic architecture.**

**a**, Comparison between the digital and optical convolution process. A random 3 × 3 kernel, normalized within [−1, 1], was defined to digitally convolve an image. The equivalent optical PSF was designed and simulated by the angular spectrum propagation method, with the optical output calculated based on the premise of a coma-free system. **b**, Architecture of the compound meta-optic forms three independent focal spots as the PSF. Angular multiplexing is used in the first layer of the metasurface, which can split light into multiple signal channels and correct the wavefront for wide-view-angle imaging. Meanwhile, polarization multiplexing is used to realize an independent response for orthogonal-polarization states. In our case, RCP and LCP signals are used for the positive and negative kernel values, respectively.

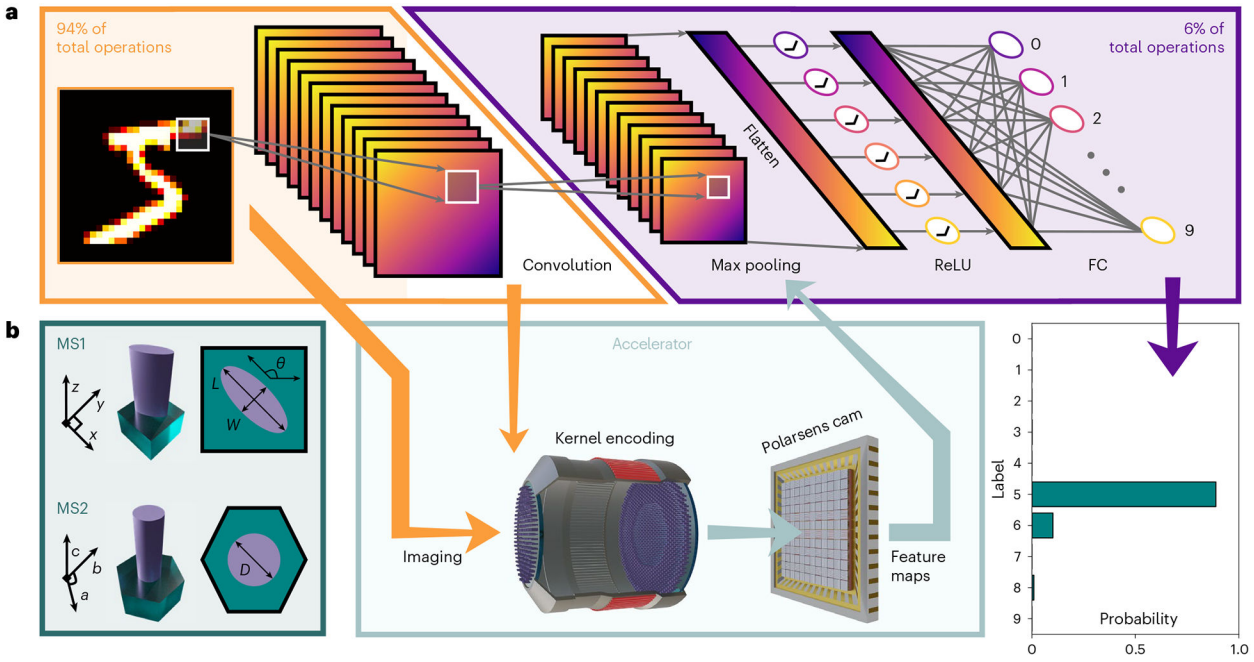


**Fig. 3 |. Design of the meta-imager.**

**a**, Design process of the hybrid neural network. A shallow convolutional neural network was trained at first. In this case, the input is convoluted by 12 independent channels, each comprising 7 × 7 pixel kernels. The convolution operations are implemented using the meta-imager, with the extracted feature maps, including multiplexed polarization channels, recorded by a polarization-sensitive camera (polarsens cam). The processed feature maps were then fed into the pretrained digital neural network to obtain the probability histogram for image classification. The percentage of relevant computing operations is indicated in the corner. ReLU, rectified linear unit; FC, fully connected. **b**, Schematic of the meta-atoms for the first (MS1) and second (MS2) metasurfaces. The height is fixed at 0.63 μm, whereas the lattice constant is chosen as 0.45 and 0.47 μm, respectively.

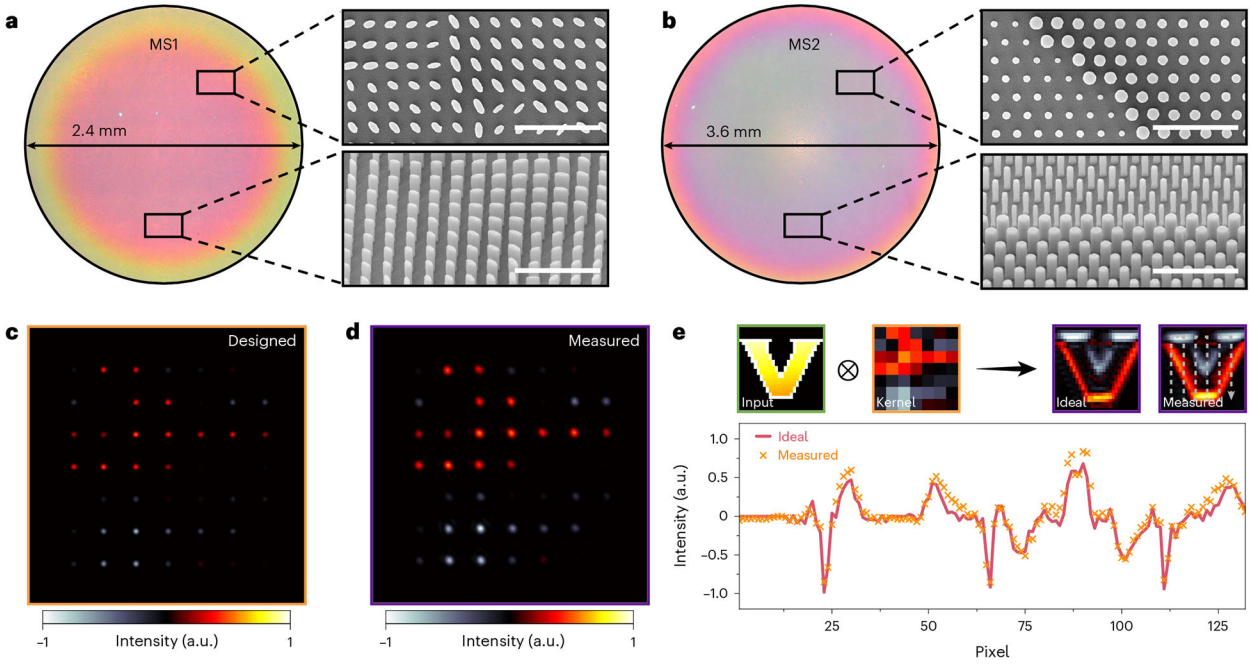


**Fig. 4 |. Fabrication and characterization of the meta-imager.**

**a**,**b**, Optical images of the fabricated metasurfaces comprising the meta-imager. The inset is a scanning electron microscopy image of each metasurface. Scale bar, 5 μm. **c**, An ideal optical kernel calculated using the angular spectrum propagation method. The weight of each spot is equal to the predesigned digital kernel. **d**, Measured intensity profile of the kernel generated by the fabricated meta-optic. **e**, Comparison between convolutional results based on the ideal and measured kernels. The dashed white line indicates the sampled pixels for comparison. The demonstration kernel is the same as those in **c** and **d**.

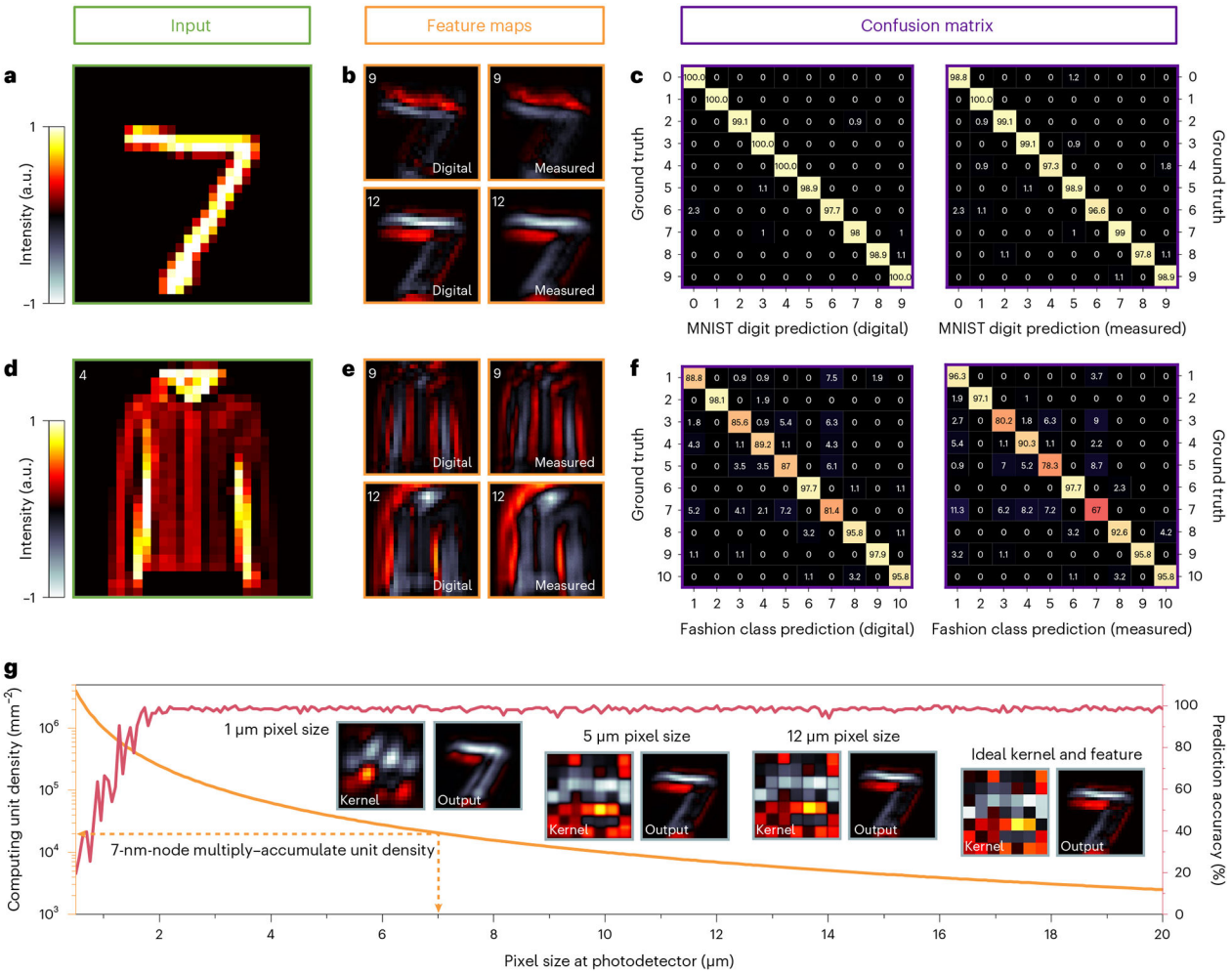


**Fig. 5 |. Classification of MNIST and Fashion MNIST objects.**

**a**, An input image from the MNIST dataset. **b**, Ideal and experimentally measured feature maps corresponding to the convolution of the data in **a** with channels 9 and 12. The top-left corner label indicates the channel number during convolution. **c**, Comparison between the theoretical and measured confusion matrices for MNIST classification. **d**, An input image from the Fashion MNIST dataset. The top-left corner label indicates the object class number. **e**, Ideal and experimentally measured feature maps corresponding to the convolution of the data in **d** with channels 9 and 12. The top-left corner label indicates the channel number during convolution. **f**, Comparison between the theoretical and measured confusion matrices for Fashion MNIST classification. **g**, Predicted accuracy curve for the MNIST dataset and the areal density of the basic computing unit as a function of pixel size. The insets depict the kernel profiles and feature maps at different pixel sizes.