

10

11 **Multi-channel Meta-imagers for Accelerating Machine Vision**

12 Hanyu Zheng¹, Quan Liu², Ivan I. Kravchenko³, Xiaomeng Zhang⁴, Yuankai Huo², and
13 Jason G. Valentine^{4*}.

- 14 1. Department of Electrical and Computer Engineering, Vanderbilt University, Nashville,
15 TN, USA, 37212.
16 2. Department of Computer Science, Vanderbilt University, Nashville, TN, USA, 37212.
17 3. Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge,
18 TN, USA, 37830.
19 4. Department of Mechanical Engineering, Vanderbilt University, Nashville, TN, USA,
20 37212.

21 * Corresponding author: jason.g.valentine@vanderbilt.edu

22 **Abstract**

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

35 Main

36 The rapid development of digital neural networks and the availability of large training
37 datasets have enabled a wide range of machine-learning-based applications, including image
38 analysis^{1, 2}, speech recognition^{3, 4}, and machine vision⁵. However, enhanced performance is
39 typically associated with a rise in model complexity, leading to larger compute requirements⁶. The
40 escalating use and complexity of neural networks have resulted in increases in energy consumption
41 while limiting real-time decision-making when large computational resources are not readily
42 accessible. These issues are especially critical to the performance of machine vision^{7, 8, 9} in
43 autonomous systems where the imager and processor must have small size, weight, and power
44 consumption for on-board processing while still maintaining low latency, high accuracy, and
45 highly robust operation. These opposing requirements necessitate the development of new
46 hardware and software solutions as the demands on machine vision systems continue to grow.

47 Optics has long been studied as a way to speed computational operations while also
48 increasing energy efficiency^{10, 11, 12, 13, 14, 15, 16}. In accelerating vision systems there is the unique
49 opportunity to off-load computation into the front-end imaging optics by designing an imager that
50 is optimized for a particular computational task. Free-space optical computation, based on
51 Fourier optics^{17, 18, 19, 20}, actually predates modern digital circuitry and allows for highly parallel
52 execution of the convolution operations which comprise the majority of the floating point
53 operations (FLOPs) in machine vision architectures^{21, 22}. The challenge with Fourier-based
54 processors is that they are traditionally employed by reprojecting the imagery using spatial light
55 modulators and coherent sources, enlarging the system size compared to chip-based approaches^{23,}
56 ^{24, 25, 26, 27, 28}. While coherent illumination is not strictly required, it allows for more freedom in the
57 convolution operations including the ability to achieve the negatively valued kernels needed for
58 spatial derivatives. Optical diffractive neural networks^{29, 30, 31} offer an alternative approach though
59 these are also employed with coherent sources and thus are best suited as back-end processors with
60 image data being reprojected.

61 Metasurfaces offer a unique platform for implementing front-end optical computation as
62 they can reduce the size of the optical elements while allowing for a wider range of optical
63 properties including polarization^{32, 33}, wavelength^{34, 35}, and angle of incidence^{36, 37} to be utilized in
64 computation. For instance, metasurfaces have been demonstrated with angle of incidence
65 dependent transfer functions for realizing compact optical differentiation systems^{38, 39, 40, 41} with
66 no need to pass through the Fourier plane of a two lens system. In addition, wavelength multiplexed
67 metasurfaces, combined with optoelectronic subtraction, have been used to achieve negatively valued
68 kernels for executing single-shot differentiation with incoherent light^{42, 43}. Differentiation,
69 however, is a single convolution operation while most machine vision systems require multiple
70 independent channels. There has been recent work on multi-channel convolutional front-ends but
71 these have been limited in transmission efficiency and computational complexity, achieving only
72 positively valued kernels with a stride that is equal to the kernel size, preventing implementation
73 of common digital designs^{44, 45}. While these are important steps towards a computational front-end,
74 an architecture is still needed for generating the multiple independent, and arbitrary, convolution
75 channels that are used in machine vision systems.

76 Here, we demonstrate a meta-imager that can serve as a multi-channel convolutional
77 accelerator for incoherent light. To achieve this, the point spread function (PSF) of the imaging
78 meta-optic is engineered to achieve parallel multi-channel convolution using a single aperture

79 implemented with angular multiplexing, as shown in Fig.1. In addition, positively and negatively
80 valued kernels are achieved for incoherent illumination by using polarization multiplexing⁴⁶,
81 combined with a polarization-sensitive camera and optoelectronic subtraction. A second
82 metasurface corrector is also employed to widen the field of view (FOV) for imaging objects in
83 the natural world and both metasurfaces are restricted to phase functions, yielding high
84 transmission efficiency. As a proof-of-concept, the platform is used to experimentally demonstrate
85 classification of the MNIST and Fashion-MNIST datasets⁴⁷ with measured accuracies of 98.6%
86 and 88.8%, respectively. In both cases, 94% of the operations are off-loaded from the digital
87 platform into the front-end optics.

88 **Angular and Polarization Multiplexing**

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

105 In this architecture, positively and negatively valued kernel weights are achieved by
106 encoding the focal spots with either right-hand-circular polarization (RCP) or left-hand-circular
107 (LCP), respectively. The circular-polarized signal is decoded by using a quarter waveplate (QWP)
108 combined with a polarization-sensitive camera containing four directional gratings integrated onto
109 each pixel. The RCP and LCP encoded feature maps, shown in Fig. 2(a), are then independently
110 recorded using the polarization-sensitive camera with summing being achieved by digitally
111 subtracting the LCP feature map from the RCP feature map. The convolution generated by this
112 method is identical to the digital process which is evidenced by comparing the digital and optical
113 feature maps in Fig. 2(a). We have used this approach for several reasons. First, as will be
114 explained, the phase and amplitude profile associated with our desired $PSF(x, y)$ is analytical,
115 significantly simplifying the design process and allowing us to achieve numerous independent
116 feature maps, or channels, using one aperture. In addition, since we have a true optical analog to a
117 digital system, we can directly implement digital kernel designs with optics, removing the optic
118 from the design loop, further speeding the design process. In order to achieve the desired optical
119 response, we employ a bilayer metasurfaces architecture, as shown in Fig. 2(b). In this architecture,
120 the first metasurface splits the incident signal into angular channels of varying weight while
121 birefringence in this layer is used to encode positive and negative kernel values in RCP and LCP

122 polarization, respectively. The second metasurface is polarization insensitive and serves as the
 123 focusing optic to create a $N \times N$ focal spot array for each channel.

124 Meta-optic Design

125 Meta-optic design began by optimizing a two metasurface lens, comprising a wavefront
 126 corrector and focuser, to be coma-free over a $\pm 10^\circ$ angular range using the commercial software,
 127 Zemax (see details in methods). The phase profiles and angular response of the metasurfaces can
 128 be found in the supplementary note S1, which shows constant focal spot shape within the designed
 129 angular range. Wider FOV can be achieved by further cascading metasurfaces as shown in
 130 supplementary note S2. Once the coma-free meta-optic was designed, angular multiplexing was
 131 applied to the first metasurface to form focal spot arrays as the convolution kernels. The focal spot
 132 position is controlled using angular multiplexing with each angle corresponding to a kernel pixel.
 133 By encoding a weight to each angular component, the system PSF, serving as the optical kernel,
 134 can be readily engineered. The analytical expression of the complex-amplitude profile
 135 multiplexing all angular signals is given by,

$$136 \quad A(x, y) = \sum_m^M \sum_n^N \sqrt{w_{mn}} \exp \left\{ i \frac{2\pi}{\lambda} [x \sin(\theta_{x|mn}) + y \sin(\theta_{y|mn})] \right\} \quad (1)$$

137 where $A(x, y)$ is a complex-amplitude field. M, N is the row and column number of elements in
 138 the kernel. w_{mn} is the corresponding weight of each element, which is normalized to a range of
 139 $[0, 1]$. λ is the working wavelength, x and y are the spatial coordinates, and $\theta_{x|mn}$ and $\theta_{y|mn}$ are
 140 the designed angles with a small variation to form the kernel elements. The deflection angles are
 141 selected to realize the desired PSF for incoherent light illumination which is given by,

$$142 \quad PSF(x, y) = \sum_m^M \sum_n^N w_{mn} \Theta \left\{ x - f_1 c \left[\frac{x_0}{f_2} + \tan(\theta_{x|mn}) \right], y - f_1 c \left[\frac{y_0}{f_2} + \tan(\theta_{y|mn}) \right] \right\} \quad (2)$$

143 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.
 144 f_1 is focal length of the meta-imager while c is a constant fitted based on the imaging system. f_2
 145 is the distance from the object to the front aperture. The detailed derivative can be found in the
 146 supplementary note S3. The separation distance of each focal spot, Δp , defines the imaged pixel
 147 size of the object. Based on a prescribed PSF the required angles, θ_{mn} , can be derived from Eq.2,
 148 which can be further extended into an off-axis imaging case, as exhibited in supplementary note
 149 S4, for the purpose of multi-channel, single-shot convolutional applications.

150 In Eq.1 we employ a spatially varying complex-valued amplitude function (see the
 151 workflow of the design process in supplementary note S5) that would ultimately introduce large
 152 reflection loss leading to a low diffraction efficiency⁴⁸. To overcome this limitation, an
 153 optimization platform was developed based on the angular spectrum propagation method and
 154 stochastic gradient descent (SGD) solver, which converts the complex-amplitude profile into a
 155 phase-only metasurface. The algorithm encodes a phase term, $\exp(i\phi_{mn})$, onto each weight, w_{mn} ,
 156 based on the loss function, $\mathcal{L} = \sum (|A|^2 - I)^2 / N$. Here, I is a matrix consisting of unity elements
 157 and N is the total pixel number. The intensity profile becomes more consistent and closer to a
 158 phase-only device by minimizing the loss function during optimization (see the detailed algorithm
 159 in supplementary note S6). The phase-only approximation can effectively avoid loss in the

160 complex-amplitude function, leading to a theoretical diffraction efficiency as high as 84.3% where
161 14% of the loss is introduced by Fresnel reflection, which can be removed by adding anti-reflection
162 coatings.

163 **Hybrid Neural Network for Object Classification**

164 In order to validate the performance of this architecture, a shallow CNN was trained for
165 the purpose of image classification. The neural network architecture, shown in Fig.3 (a), contains
166 an optical convolution layer followed by digital max pooling, a rectified linear (ReLU) activation
167 function, and a fully connected (FC) layer. In the convolution process, 12 independent kernels are
168 used to extract feature maps and the overall intensity of positive and negative channels was set to
169 be equal due to energy conservation from the phase-only approximation in the meta-optic design.
170 Since neural network training is a high-dimensional problem with infinite solutions, the above
171 kernel restrictions do not significantly affect the final performance (see discussion in the
172 supplementary note S7). Each kernel comprised $N = 7$ pixels instead of a more typical $N = 3$ format,
173 to correlate neurons within a broader viewing field⁴⁹, leading to better performance for large-scale
174 object recognition. The detailed training process is described in the methods section. In order to
175 finish classification, the feature maps extracted by the compound meta-optic are fed into the digital
176 component of the neural network. In this architecture 94% of the total operations are off-loaded
177 from the digital platform into the meta-optic leading to a significant speedup for classification
178 tasks (see the details in supplementary note S8).

179 **Meta-optic Implementation**

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

$$186 \begin{bmatrix} \phi_{LCP} \\ \phi_{RCP} \end{bmatrix} = \begin{bmatrix} \phi_x + 2\theta + \pi/4 \\ \phi_x - 2\theta - \pi/4 \end{bmatrix} \quad (3)$$

187 Here, ϕ_x is the phase delay of the meta-atoms along x axis at $\theta = 0$. Hence, by tuning the length,
188 width, and rotation angle, the phase delay of LCP and RCP light can be independently controlled
189 (see the detailed derivative process in supplementary note S9). The second metasurface was
190 designed based on circular nanopillars arranged in a hexagonal lattice for realizing polarization-
191 insensitive phase control. The phase delay of the circular nanopillars as a function of diameter can
192 be found in supplementary note S10.

193 **Fabrication and Characterization of Meta-optic**

194 Two versions of the meta-optic classifier were fabricated based on networks trained for
195 MNIST and Fashion-MNIST datasets, with one set of the phase profiles shown in supplementary
196 note S11. Fabrication of the meta-optic began with a silicon device layer on a fused silica substrate
197 patterned by the standard electron beam lithography (EBL) followed by reactive-ion-etching (RIE).

198 A thin polymethyl methacrylate (PMMA) layer was spin-coated over the device as the protective
199 and index-matching layer. The detailed fabrication process is described in the Methods section.
200 An optical image of the two metasurfaces comprising the meta-optic is exhibited in Fig.4 (a) and
201 (b) with the inset showing the meta-atoms. In order to align the compound meta-optic, the first
202 metasurface was mounted in a rotational stage (CRM1PT, Thorlabs) while the second layer was
203 fitted in a 3-axis translational stage (CXYZ05A, Thorlabs). The metasurfaces are aligned in situ
204 and characterized in a cage system, with the detailed alignment setup shown in supplementary note
205 S12. A meta-hologram was fabricated on the first layer alongside the device to assist the alignment
206 process by forming an alignment pattern at a prescribed distance along the optical axis
207 corresponding to the designed separation distance. The alignment process was finished by
208 overlapping the alignment pattern with the low-transmission register on the second layer. Due to
209 the large size (mm-scale) of each metasurface layer, the meta-optic exhibits high alignment
210 tolerance. The system performance remains constant under a horizontal misalignment of $65\mu\text{m}$
211 and vertical displacement of $\pm 400\mu\text{m}$, indicating the robustness of the entire convolutional system.
212 The alignment error analysis can be found in supplementary note S13.

213 In order to characterize the optical properties of the fabricated meta-optic, a linearly
214 polarized laser was used for illumination in obtaining the PSF (see the detailed characterization
215 setup in supplementary note S14). The linearly polarized light source includes LCP and RCP
216 components with equal strength. The PSF at the focal plane of the compound meta-optic, shown
217 in Fig.4 (c) and (d), indicates a good match between the ideal and measured results, where the red
218 and blue represent positive and negative values, respectively.

219 Optical convolution of a grayscale Vanderbilt logo was used to characterize the accuracy
220 of the fabricated meta-optic, as shown in Fig.4 (e). To accomplish this, an imaging system using a
221 liquid-crystal-based spatial light modulator (SLM) was built with the details shown in
222 supplementary note S15. An incoherent tungsten lamp with a 10 nm wide bandpass filter was used
223 for SLM illumination. The feature maps extracted by the meta-optic were recorded by a
224 polarization-sensitive camera (DZK 33UX250, Imaging Source) where orthogonally polarized
225 channels are simultaneously recorded using polarization filters on each camera pixel. The
226 comparison between the digital and measured feature maps, recorded on the camera, is illustrated
227 in Fig.4 (e). The pixel intensity from digital and measured convolutional results at the same
228 position were extracted and compared to evaluate the convolution fidelity. The deviation between
229 the ideal and measured results, defined by $\sigma = \sum_{n=1}^N |D_{i,n} - D_{m,n}| / (2N)$, was calculated as 3.83%,
230 where D_i and D_m are the ideal and measured intensity, and N is the number of total pixels. The
231 error originates from stray light, fabrication imperfections, the local phase approximation, and
232 metasurface misalignment (see the detailed system error analysis in supplementary note S16).
233 These errors also result in a small amount of zeroth order diffracted light being introduced from
234 the first metasurface leading to a spot at the center of the imaging plane. However, the polarization
235 state of the zeroth order light remains unchanged, with the energy evenly distributed in the two
236 circular polarized channels. Hence, subtraction between the information channels allows the zeroth
237 order pattern to be canceled, not affecting the classification performance. The detailed discussion
238 can be found in supplementary note S17.

239 **Object Classification for Machine Vision**

240 As a proof-of-concept in demonstrating multi-channel convolution, a full meta-optic
241 classifier was first designed and fabricated based on classification of the MNIST dataset, which
242 includes 60,000 hand-written digit training images with 28×28 pixel format. The feature maps
243 of 1000 digits, not in the training set, were extracted using the meta-optic to characterize the system
244 performance. An example input image is exhibited in Fig.5 (a), with the corresponding feature
245 maps shown in Fig.5 (b). The kernels and feature maps for all the channels are illustrated in
246 supplementary note S18. The measured feature maps match well with the theoretical prediction,
247 as shown in Fig.5 (b), indicating good fidelity in the optical convolution process. The theoretical
248 and experimental confusion matrices for this testing dataset are shown in Fig.5 (c), demonstrating
249 99.3% accurate classification in theory and 98.6% accurate classification in the measurement. The
250 small drop in accuracy likely results from the small inaccuracy in the realized optical kernels.
251 While the system was designed at a single wavelength, simulations indicate minimal accuracy drop
252 up to an illumination bandwidth of 50 nm, indicating that the experimental bandwidth of 10 nm
253 should have a minimal impact (see detailed discussion in supplementary note S19).

254 In order to explore the flexibility of the approach a dataset with higher spatial frequency
255 information, Fashion-MNIST, was also used for training the model with an example input image
256 provided in Fig.5 (d). This dataset includes 60,000 training images of clothing articles that contain
257 images with higher spatial frequencies than the MNIST handwritten digit dataset. The ideal and
258 measured feature maps are compared in Fig.5 (e), indicating good agreement. All of the designed
259 kernel profiles and feature maps are shown in supplementary note S20. The confusion matrices for
260 Fashion-MNIST are illustrated in Fig.5 (f), with 90.2% accurate classification in theory and 88.8%
261 in measurement. To validate the significance of the optical convolution layer, a reference model
262 for MNIST handwritten digit classification, without a convolutional layer, was trained, resulting
263 in an accuracy of 80.3%, illustrating the importance of the convolution operations (see detailed
264 discussion in supplementary note S21). Compared to the MNIST dataset, the Fashion-MNIST
265 model has a slightly lower accuracy, in theory, due to the higher resolution features in the dataset.
266 Specifically, for class 7 in the Fashion-MNIST dataset, the accuracy predicted by the optical
267 frontend dropped from 81.4% to 67.0%, with the model miss-identifying the images as classes
268 1,3,4,5. We expect these classes to share the same features during model training (see discussion
269 in the supplementary note S22). These mixed features can be potentially distinguished by
270 adaptively tuning the loss function during model training⁵⁰ or utilizing novel neural network
271 architecture such as vision transformer⁵¹ (ViT) with better performance at comparable FLOPs.

272 To understand the scalability of the meta-imager, the accuracy of classification as a
273 function of the areal density of the basic computing unit was calculated, as shown in Fig.5 (g). The
274 optical computing unit density is defined as the convolutional pixels per unit area where we assume
275 each convolutional pixel is matched to a physical pixel on a photodetector. The pixel size is
276 dictated by the separation distance between the neighboring focal spots in the PSF, which is
277 ultimately dictated by the diffraction limit. The prediction accuracy is based on the MNIST dataset
278 and the theoretical accuracy remains as high as $\sim 99\%$ until the pixel size drops below $2\mu\text{m}$, at
279 which point neighboring focal spots are below the diffraction limit, resulting in additional
280 aberration in the output features, as shown in the inset images in Fig.5 (g). Thus, although a pixel
281 size of $12\mu\text{m}$ is demonstrated in this work as a proof of concept, the system functionality would
282 remain unchanged, in theory, with up to 6X higher areal computing unit density. For perspective,
283 the meta-imager computing unit density can be compared to the multiply-accumulation (MAC)

284 unit density and size based on the current 7nm node architecture⁵², which results in MACs with a
285 size of $\sim 7\mu\text{m} \times 7\mu\text{m}$.

286 **Conclusions**

287 Our meta-imager is a proof-of-concept for a convolutional front-end that can be used to
288 replace the traditional imaging optics in machine vision applications, encoding information in a
289 more efficient basis for back-end processing. In this context, negatively valued kernels and multi-
290 channel convolution, enabled by meta-optics, allows one to increase the number of operations that
291 can be off-loaded into the front-end optics. Furthermore, the architecture allows for incoherent
292 illumination and a reasonably wide FOV, both of which are needed for implementation in imaging
293 natural scenes with ambient illumination. Although a tradeoff exists between the channel number
294 and the viewing angle range, a multi-aperture architecture could be designed without deteriorating
295 the FOV in a single imaging channel⁵³. In addition, we have not attempted to optimize the
296 operation bandwidth, which could be addressed through dispersion engineering, over modest
297 apertures, combination with broadband refractive optics, or use of dispersion to perform
298 wavelength-dependent functions. Further acceleration can be realized via integration of a meta-
299 imager front-end directly with a chip-based photonics back-end such that data readout and
300 transport can be achieved without analog-to-digital converters for ultrafast and low-latency
301 processing.

302 Our meta-imager does put restrictions on the depth, or number of layers, in the optical
303 front-end which means that it provides the most benefit in lightweight neural networks such as
304 those found in power-limited or high-speed autonomous applications. Recent advances in machine
305 learning, such as the use of larger kernels for network layer compression⁵⁴ and re-
306 parameterization⁵⁵ could further improve the effectiveness of single, or few layer, meta-imager
307 front-ends. In addition, the capability of meta-optics for multi-functional processing, including
308 wavelength and polarization-based discrimination, can be used to further increase information
309 collection⁴⁴. As a result, this general architecture for meta-imagers can be highly parallel and
310 bridge the gap between the natural world and digital systems, potentially finding use beyond
311 machine vision⁵⁶ in applications such as information security^{57,58} and quantum communications⁵⁹.

312 **Acknowledgments**

313 HZ and JGV acknowledge support from DARPA under contract HR001118C0015 and
314 NAVAIR under contract N6893622C0030. XZ acknowledges support from ONR under contract
315 N000142112468. YH and QL acknowledge support from NIH under contract R01DK135597.
316 Meta-optic devices were manufactured as part of a user project at the Center for Nanophase
317 Materials Sciences (CNMS), which is a US Department of Energy, Office of Science User Facility
318 at Oak Ridge National Laboratory.

319 **Author Contributions**

320 HZ and JGV developed the idea. HZ conducted the optical modeling and system design.
321 QL and HZ trained the digital neural network. HZ fabricated the samples. IIK performed silicon

322 growth and EBL for the metasurfaces. HZ conducted the experimental measurements. HZ, QL,
323 and XZ executed the data analysis. HZ and JGV wrote the manuscript with input from all the
324 authors. The project was supervised by YH and JGV.

325 **Competing Interest**

326 The authors declare no competing interests.

327 **Figure Legends/Captions**

328 *Figure 1. Schematic of the meta-imager.* The meta-imager enables multi-channel signal processing
329 for replacing convolution operations in a digital neural network. A bilayer meta-optic system
330 encoded by the pre-designed kernels is utilized to achieve optical convolution with the incoherent
331 light source to be used for object illumination. Positive and negative values are distinguished and
332 recorded as feature maps by a polarization-sensitive photodetector, where an oriented grating sits
333 on each photodetector pixel for polarized signal sorting.

334 *Figure 2. Meta-optic Architecture.* (a) Comparison between the digital and optical convolution
335 process. A random 3×3 kernel, normalized between $[-1,1]$, was defined to convolve an image
336 digitally. The equivalent optical PSF was designed and simulated by the angular spectrum
337 propagation method, with the optical output calculated based on the premise of a coma-free system.
338 (b) The architecture of the compound meta-optic forms three independent focal spots as the PSF.
339 Angular multiplexing is used in the first layer metasurface, which can split light into multiple
340 signal channels and correct the wavefront for wide-view-angle imaging. Meanwhile, polarization
341 multiplexing is used to realize an independent response for orthogonal polarization states. In our
342 case right-hand-circular (RCP) and left-hand-circular (LCP) polarized signals are used for positive
343 and negative kernel values, respectively.

344 *Figure 3. Design of the Meta-imager.* (a) Design process of the hybrid neural network. A shallow
345 convolutional neural network was trained at first. In this case, the input is convoluted by 12
346 independent channels, each comprising 7×7 pixel kernels. The convolution operations are
347 implemented using the meta-imager, with the extracted feature maps, including multiplexed
348 polarization channels, recorded by a polarization-sensitive camera. The processed feature maps
349 were then fed into the pre-trained digital neural network to obtain the probability histogram for
350 image classification. The number at the corner indicates the percentage of relevant computing
351 operations. (b) The schematic of the meta-atoms for the first and second metasurfaces. The height
352 is fixed at $0.6\mu\text{m}$ while the lattice constant is chosen as $0.45\mu\text{m}$ and $0.47\mu\text{m}$, respectively.

353 *Figure 4. Fabrication and Characterization of the Meta-imager.* (a) and (b) Optical images of the
354 fabricated metasurfaces comprising the meta-imager. The inset is an SEM image of each
355 metasurface. Scale bar: $5\mu\text{m}$. (c) An ideal optical kernel calculated based on the angular spectrum
356 propagation method. The weight of each spot is equal to the pre-designed digital kernel. (d)
357 Measured intensity profile of the kernel generated by the fabricated meta-optic. (e) Comparison
358 between convolutional results based on the ideal and measured kernels. The solid white line
359 indicates the sampled pixels for comparison. The demonstration kernel is the same as (c) and (d).

360 *Figure 5. Classification of MNIST and Fashion-MNIST objects.* (a) An input image from the
361 MNIST dataset. (b) Ideal and experimentally measured feature maps corresponding to the
362 convolution of (a) with channels 1 and 4. The upper-left corner label indicates the channel number
363 during convolution. (c) Comparison between the theoretical and measured confusion matrices for
364 MNIST classification. (d) An input image from the Fashion-MNIST dataset (e) Ideal and
365 experimentally measured feature maps corresponding to the convolution of (a) with channels 1
366 and 4. The upper-left corner label indicates the channel number during convolution. (f)
367 Comparison between the theoretical and measured confusion matrices for Fashion-MNIST
368 classification. (g) Predicted accuracy curve for the MNIST dataset and the areal density of basic
369 computing unit as a function of pixel size. The insets depict kernel profiles and feature maps at
370 different pixel sizes.

371 **References**

- 372 1. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image
373 recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference*
374 *Track Proceedings* 1–14 (2015).
- 375 2. Wang, G. *et al.* Interactive Medical Image Segmentation Using Deep Learning with Image-
376 Specific Fine Tuning. *IEEE Trans Med Imaging* **37**, 1562–1573 (2018).
- 377 3. Furui, S., Deng, L., Gales, M., Ney, H. & Tokuda, K. Fundamental technologies in modern
378 speech recognition. *IEEE Signal Process Mag* **29**, 16–17 (2012).
- 379 4. Sak, H., Senior, A., Rao, K. & Beaufays, F. Fast and accurate recurrent neural network
380 acoustic models for speech recognition. *Proceedings of the Annual Conference of the International*
381 *Speech Communication Association, INTERSPEECH 2015-Janua*, 1468–1472 (2015).
- 382 5. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition.
383 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*
384 *Recognition 2016-Decem*, 770–778 (2016).
- 385 6. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- 386 7. Menzel, L. *et al.* Ultrafast machine vision with 2D material neural network image sensors.
387 *Nature* **579**, 62–66 (2020).
- 388 8. Liu, L. *et al.* Computing Systems for Autonomous Driving: State of the Art and Challenges.
389 *IEEE Internet Things J* **8**, 6469–6486 (2021).
- 390 9. Shi, W. *et al.* LOEN: Lensless opto-electronic neural network empowered machine vision.
391 *Light Sci Appl* **11**, (2022).
- 392 10. Hamerly, R., Bernstein, L., Sludds, A., Soljačić, M. & Englund, D. Large-Scale Optical
393 Neural Networks Based on Photoelectric Multiplication. *Phys Rev X* **9**, 1–12 (2019).
- 394 11. Wetzstein, G. *et al.* Inference in artificial intelligence with deep optics and photonics.
395 *Nature* **588**, 39–47 (2020).

- 396 12. Shastri, B. J. *et al.* Photonics for artificial intelligence and neuromorphic computing. *Nat*
397 *Photonics* **15**, 102–114 (2021).
- 398 13. Xue, W. & Miller, O. D. High-NA optical edge detection via optimized multilayer films.
399 *Journal of Optics (United Kingdom)* **23**, (2021).
- 400 14. Wang, T. *et al.* An optical neural network using less than 1 photon per multiplication. *Nat*
401 *Commun* **13**, 123 (2022).
- 402 15. Wang, T. *et al.* Image sensing with multilayer nonlinear optical neural networks. *Nat*
403 *Photonics* **17**, 8–17 (2023).
- 404 16. Badloe, T., Lee, S. & Rho, J. Computation at the speed of light: metamaterials for all-
405 optical calculations and neural networks. *Advanced Photonics* vol. 4 Preprint at
406 <https://doi.org/10.1117/1.AP.4.6.064002> (2022).
- 407 17. Vanderlugt, A. Optical signal processing. *Wiley* (1993).
- 408 18. Chang, J., Sitzmann, V., Dun, X., Heidrich, W. & Wetzstein, G. Hybrid optical-electronic
409 convolutional neural networks with optimized diffractive optics for image classification. *Sci Rep*
410 **8**, 1–10 (2018).
- 411 19. Colburn, S., Chu, Y., Shilzerman, E. & Majumdar, A. Optical frontend for a convolutional
412 neural network. *Appl Opt* **58**, 3179 (2019).
- 413 20. Zhou, T. *et al.* Large-scale neuromorphic optoelectronic computing with a reconfigurable
414 diffractive processing unit. *Nat Photonics* **15**, 367–373 (2021).
- 415 21. Chen, Y. H., Krishna, T., Emer, J. S. & Sze, V. Eyeriss: An Energy-Efficient
416 Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE J Solid-State Circuits*
417 **52**, 127–138 (2017).
- 418 22. Neshatpour, K., Homayoun, H. & Sasan, A. ICNN: The iterative convolutional neural
419 network. *ACM Transactions on Embedded Computing Systems* **18**, (2019).
- 420 23. Xu, X. *et al.* 11 TOPS photonic convolutional accelerator for optical neural networks.
421 *Nature* **589**, 44–51 (2021).
- 422 24. Feldmann, J. *et al.* Parallel convolutional processing using an integrated photonic tensor
423 core. *Nature* **589**, 52–58 (2021).
- 424 25. Wu, C. *et al.* Programmable phase-change metasurfaces on waveguides for multimode
425 photonic convolutional neural network. *Nat Commun* **12**, 1–8 (2021).
- 426 26. Zhang, H. *et al.* An optical neural chip for implementing complex-valued neural network.
427 *Nat Commun* **12**, 1–11 (2021).
- 428 27. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for
429 image classification. *Nature* **606**, 501–506 (2022).

- 430 28. Fu, T. *et al.* Photonic machine learning with on-chip diffractive optics. *Nat Commun* **14**,
431 1–10 (2023).
- 432 29. Lin, X. *et al.* All-optical machine learning using diffractive deep neural networks. *Science*
433 (1979) **361**, 1004–1008 (2018).
- 434 30. Qian, C. *et al.* Performing optical logic operations by a diffractive neural network. *Light*
435 *Sci Appl* **9**, (2020).
- 436 31. Luo, X. *et al.* Metasurface-enabled on-chip multiplexed diffractive neural networks in the
437 visible. *Light Sci Appl* **11**, (2022).
- 438 32. Kwon, H., Arbabi, E., Kamali, S. M., Faraji-Dana, M. S. & Faraon, A. Single-shot
439 quantitative phase gradient microscopy using a system of multifunctional metasurfaces. *Nat*
440 *Photonics* **14**, 109–114 (2020).
- 441 33. Xiong, B. *et al.* Breaking the limitation of polarization multiplexing in optical metasurfaces
442 with engineered noise. *Science* **379**, 294–299 (2023).
- 443 34. Khorasaninejad, M. *et al.* Metalenses at visible wavelengths: Diffraction-limited focusing
444 and subwavelength resolution imaging. *Science (1979)* **352**, 1190–1194 (2016).
- 445 35. Kim, J. *et al.* Scalable manufacturing of high-index atomic layer–polymer hybrid
446 metasurfaces for metapotonics in the visible. *Nat Mater* **22**, 474–481 (2023).
- 447 36. Levanon, N. *et al.* Angular Transmission Response of In-Plane Symmetry-Breaking Quasi-
448 BIC All-Dielectric Metasurfaces. *ACS Photonics* **9**, 3642–3648 (2022).
- 449 37. Nolen, J. R., Overvig, A. C., Cotrufo, M. & Alù, A. Arbitrarily polarized and unidirectional
450 emission from thermal metasurfaces. *arXiv* (2023).
- 451 38. Guo, C., Xiao, M., Minkov, M., Shi, Y. & Fan, S. Photonic crystal slab Laplace operator
452 for image differentiation. *Optica* **5**, 251 (2018).
- 453 39. Cordaro, A. *et al.* High-Index Dielectric Metasurfaces Performing Mathematical
454 Operations. *Nano Lett* **19**, 8418–8423 (2019).
- 455 40. Zhou, Y., Zheng, H., Kravchenko, I. I. & Valentine, J. Flat optics for image differentiation.
456 *Nat Photonics* **14**, 316–323 (2020).
- 457 41. Fu, W. *et al.* Ultracompact meta-imagers for arbitrary all-optical convolution. *Light Sci*
458 *Appl* **11**, (2022).
- 459 42. Wang, H., Guo, C., Zhao, Z. & Fan, S. Compact Incoherent Image Differentiation with
460 Nanophotonic Structures. *ACS Photonics* **7**, 338–343 (2020).
- 461 43. Zhang, X., Bai, B., Sun, H. B., Jin, G. & Valentine, J. Incoherent Optoelectronic
462 Differentiation Based on Optimized Multilayer Films. *Laser Photon Rev* **16**, 1–8 (2022).
- 463 44. Zheng, H. *et al.* Meta-optic accelerators for object classifiers. *Sci Adv* **8**, 1–9 (2022).

- 464 45. Bernstein, L. *et al.* Single-Shot Optical Neural Network. *Sci Adv* **9**, 1–10 (2023).
- 465 46. Shen, Z. *et al.* Monocular metasurface camera for passive single-shot 4D imaging. *Nat*
466 *Commun* **14**, 1–8 (2023).
- 467 47. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to
468 document recognition. *Proceedings of the IEEE* **86**, 2278–2323 (1998).
- 469 48. Zheng, H. *et al.* Compound Meta-Optics for Complete and Loss-Less Field Control. *ACS*
470 *Nano* **16**, 15100–15107 (2022).
- 471 49. Liu, S. *et al.* More ConvNets in the 2020s: Scaling up Kernels Beyond 51x51 using
472 Sparsity. *arXiv* (2022).
- 473 50. Barron, J. T. A general and adaptive robust loss function. *Proceedings of the IEEE*
474 *Computer Society Conference on Computer Vision and Pattern Recognition* **2019-June**, 4326–
475 4334 (2019).
- 476 51. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image
477 Recognition at Scale. *arXiv* (2020).
- 478 52. Stillmaker, A. & Baas, B. Scaling equations for the accurate prediction of CMOS device
479 performance from 180 nm to 7 nm. *Integration* **58**, 74–81 (2017).
- 480 53. McClung, A., Samudrala, S., Torfeh, M., Mansouree, M. & Arbabi, A. Snapshot spectral
481 imaging with parallel metasystems. *Sci Adv* **6**, 1–9 (2020).
- 482 54. Ding, X., Zhang, X., Han, J. & Ding, G. Scaling Up Your Kernels to 31×31: Revisiting
483 Large Kernel Design in CNNs. in *Proceedings of the IEEE Computer Society Conference on*
484 *Computer Vision and Pattern Recognition* vols 2022-June 11953–11965 (IEEE Computer Society,
485 2022).
- 486 55. Ding, X. *et al.* RepVgg: Making VGG-style ConvNets Great Again. in *Proceedings of the*
487 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 13728–13737
488 (IEEE Computer Society, 2021). doi:10.1109/CVPR46437.2021.01352.
- 489 56. Li, L. *et al.* Intelligent metasurface imager and recognizer. *Light Sci Appl* **8**, (2019).
- 490 57. Zhao, R. *et al.* Multichannel vectorial holographic display and encryption. *Light Sci Appl*
491 **7**, (2018).
- 492 58. Kim, I. *et al.* Pixelated bifunctional metasurface-driven dynamic vectorial holographic
493 color prints for photonic security platform. *Nat Commun* **12**, 1–9 (2021).
- 494 59. Li, L. *et al.* Metalens-array-based high-dimensional and multiphoton quantum source.
495 *Science (1979)* **368**, 1487–1490 (2020).

496

497 **Methods**

498 *Optimization of Coma-free Meta-optic.* The coma-free meta-optic contains two
499 metasurfaces, whose phase profiles were optimized by the ray tracing technique using commercial
500 optical design software (Zemax OpticStudio, Zemax LLC). The phase profile of each layer was
501 defined by even order polynomials according to the radial coordinate, ρ , as follows:

502
$$\phi(\rho) = \sum_{n=1}^5 a_n \left(\frac{\rho}{R}\right)^{2n} \quad (1)$$

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

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

521 *Numerical Simulation.* The complex transmission coefficients of the silicon nanopillars
522 were calculated using an open-source rigorous coupled wave analysis (RCWA) solver, Reticolo⁶⁰.
523 A square lattice with a period of $0.45\mu\text{m}$ was used for the first metasurface with the working
524 wavelength at $0.87\mu\text{m}$. The second metasurface was assigned a hexagonal lattice with a period of
525 $0.47\mu\text{m}$. During full-wave simulation, the index of silicon and fused silica characterized by
526 ellipsometry was set at 3.74 and 1.45, respectively.

527 *Metasurface Fabrication.* EBL-based lithography was used to fabricate all the metasurface
528 layers. First, low-pressure chemical vapor deposition (LPCVD) was utilized to deposit a 630nm
529 thick silicon device layer on a fused silica substrate. PMMA photoresist was then spin-coated on
530 the silicon layer, followed by thermal evaporation of a 10nm thick Cr conduction layer. The EBL
531 system then exposed the photoresist, and after removing the Cr layer, the pattern was developed
532 by the MIBK/IPA solution. A 30nm Al_2O_3 hard mask was deposited via electron beam evaporation,
533 followed by a lift-off process with N-methyl-2-pyrrolidone (NMP) solution. The silicon was then
534 patterned using reactive ion etching, and a $1\mu\text{m}$ thick layer of PMMA was spin-coated to encase
535 the nanopillar structures as a protective and index-matching layer.

536

537 **Data Availability**

538 The data that support the findings of this study are present in the paper, Supplementary
539 Materials and/or are available from the corresponding author upon reasonable request.

540 **Methods-only References**

541 60. Hugonin, A. J. P. & Lalanne, P. RETICOLO CODE 1D for the diffraction by stacks of
542 lamellar 1D gratings. *arXiv* (2012).









