nature communications

Article

https://doi.org/10.1038/s41467-024-55558-3

Spectral convolutional neural network chip for in-sensor edge computing of incoherent natural light

Received: 17 July 2024

Accepted: 17 December 2024

Published online: 02 January 2025

Check for updates

Kaiyu Cui[1,3] ✉, Shijie Rao[1,3], Sheng Xu[1], Yidong Huang[1] ✉, Xusheng Cai[2], Zhilei Huang[2], Yu Wang[2], Xue Feng[1], Fang Liu[1], Wei Zhang[1], Yali Li[1] & Shengjin Wang[1]

Optical neural networks are considered next-generation physical implementations of artificial neural networks, but their capabilities are limited by on-chip integration scale and requirement for coherent light sources. This study proposes a spectral convolutional neural network (SCNN) with matter meta-imaging. The optical convolutional layer is implemented by integrating very large-scale and pixel-aligned spectral filters on CMOS image sensor. It facilitates highly parallel spectral vector-inner products of incident incoherent natural light i.e., the direct information carrier, which empowers in-sensor optical analog computing at extremely high energy efficiency. To the best of our knowledge, this is the first integrated optical computing utilizing natural light. We employ the same SCNN chip for completely different real-world complex tasks and achieve accuracies of over 96% for pathological diagnosis and almost 100% for face anti-spoofing at video rates. These results indicate a feasible and scalable in-sensor edge computing chip of natural light for various portable terminals.

Artificial neural networks (ANNs) have demonstrated powerful abilities across numerous applications, such as the burgeoning ChatGPT[1] and AIGC[2], and have altered many aspects of modern society. Because vision is the most important method for both humans and machines to perceive the world, among different ANNs, convolutional neural networks (CNNs) inspired by biological vision for image processing have become one of the most commonly used ANN architectures[3]. Owing to the convolutional layers that enable CNNs to extract high-level features from raw image data and significantly reduce parametric complexity[3,4], CNNs have achieved considerable success in image recognition[5], segmentation[6], and detection[7] tasks. However, the convolutional processing of the network dominates the processing time and computing power. This leads to significant computing cost challenges and severe limitations for CNNs on leading high performance electronic computing platforms, such as graphic process units (GPU), as reflected by Moore's law[8]. The huge computational cost severely limits the deployment of CNNs on portable terminals for edge computing.

Optical neural networks (ONNs), or optical neuromorphic hardware accelerators, have been regarded as one of the most promising next-generation parallel-computing platforms to address the limitations of electronic computing, with the distinct advantages of fast computational speed, high parallelism, and low power consumption[9–14]. Existing works on ONNs have achieved fully connected neural networks (FCNs) based on the Reck design[15–19] or diffractive deep neural network (D^2NN)[20–25] and optical CNNs (OCNNs) or optical convolutional accelerators by further introducing wavelength division multiplexing[26–29], attaining extraordinary computing speed with low power consumption. However, existing on-chip OCNNs hardly accept broadband incoherent natural light, i.e., the direct information carrier. The requirement for a coherent light source limits the scale of optical matrix multiplication[30] and is insufficient for two-

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China. [2]Beijing Seetrum Technology Co., Beijing, China. [3]These authors contributed equally: Kaiyu Cui, Shijie Rao. ✉e-mail: kaiyucui@tsinghua.edu.cn; yidonghuang@tsinghua.edu.cn

dimensional (2D) convolution calculations. Moreover, in these works[11,24,26,27,31–33], broadband incoherent natural light is usually captured by digital cameras and then encoded to coherent light for optical computing (Fig. 1a), which not only degrades the energy efficiency but also loses the light field features containing rich matter information, such as spectrum, polarization and incident angle. Especially, the spectral features that can identify the composition of matter for complex vision tasks cannot be directly introduced into OCNNs.

In this work, we propose and demonstrate a spectral convolutional neural network (SCNN) based on an optoelectronic computing framework that accepts broadband incoherent natural light directly as input (Fig. 1b). Hybrid optoelectronic computing hardware with an optical convolutional layer (OCL) and a reconfigurable electrical backend is employed to leverage optical superiority without sacrificing the flexibility of digital electronics[19,20,28–30,32,34–36]. The proposed OCL works as the input and the first convolutional layer, which is implemented by integrating very large-scale, pixel-aligned integration of spectral filters on a CMOS image sensor (CIS), as is shown in Fig. 1c, d. Here, the spectral filters can utilize dispersive nanostructures or material with spectral modulation abilities. In this work, we provide two implementations of the spectral filters. The first one is based on metasurfaces which provide better spectral modulation capabilities (Fig. 1c). The second one is achieved by pigments with mass production on a 12-inch wafer (Fig. 1d). The weights of the OCL are encoded on the transmission responses of the spectral filters. It should be noted that the proposed system actually functions as a high-speed customizable hyperspectral imaging method based on the new design concepts and system framework of SCNN. However, previous

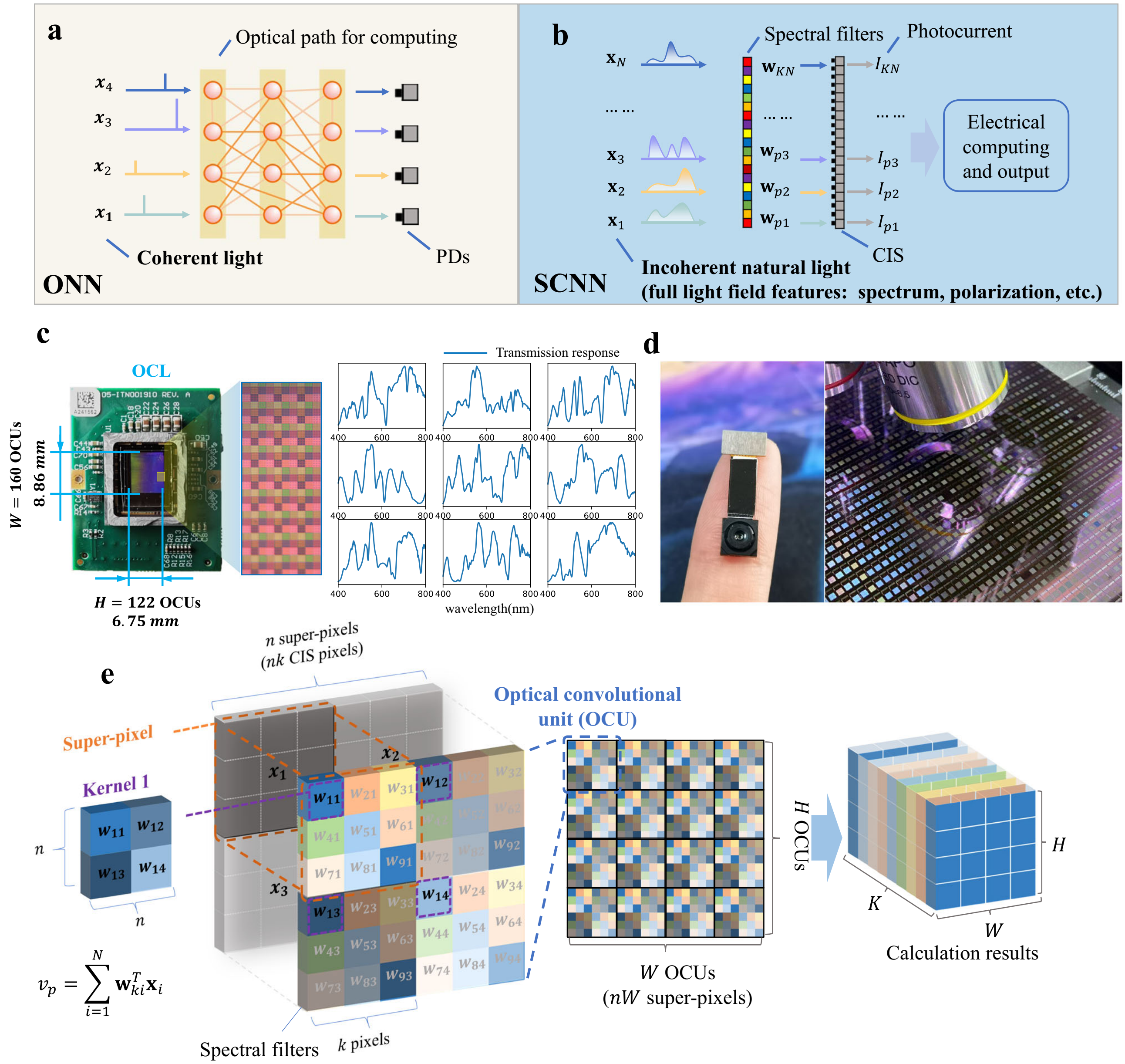


Fig. 1 | Principles of the proposed spectral convolutional neural network (SCNN). a Existing optical neural networks (ONNs) are based on coherent light sources for computing. They are incapable of broadband light field sensing and in-sensor computing. **b** In our design, we implemented an SCNN by integrating very large-scale spectral filters on CMOS image sensor (CIS). Our SCNN can accept incoherent natural light and perform analog 2D convolution calculations directly. **c** The metasurface-based optical convolutional layer (OCL) integrates pixel-aligned metasurface units on a CIS. **d** The pigment-based OCL is fabricated by lithography on a 12-inch wafer. **e** The working principles of our OCL. One OCL contains an $H \times W$ array of identical OCUs and each OCU has K convolutional kernels, resulting in calculation results of size $H \times W \times K$. $\mathbf{x}_N$: The input spectral signal. $\mathbf{w}_{KN}$: Transmission response of the spectral filter. I_{KN}: Photocurrent of the CIS pixel.

hyperspectral imaging works adopted spectral filters as the sensing matrix and got the compressively sensed hyperspectral images[37–40]. After capturing, the hyperspectral images require post-processing of spectral reconstruction and further spectral analysis. In these systems, the spectral filters are designed to achieve high spectral resolution and the post-processing of the captured data requires huge computational cost, which is incapable of applying on edge computing. In this work, the spectral filters are designed to be the first layer of the neural network. Their transmission responses work as weights of the layer rather than the sensing matrix. Therefore, we only need very few tailored spectral filters to achieve real-world applications at high efficiency because accurate spectral reconstruction is not required thus achieving edge computing. In this work, only 9 different spectral filters are designed for the SCNN. More detailed comparison is described in Supplementary Note 1.

After natural light transmits through the broadband spectral filters, CIS is used to detect the light intensity at different spatial locations (Fig. 1b), which sums the energy of the transmitted light along the wavelength axis (i.e., the spectrum) at each image pixel, similar to the functions of cone cells in the human eye. Therefore, the CIS and spectral filters form an analog OCL with high spatial resolution and process natural images directly without explicit image duplication. As the OCL facilitates a highly parallel vector inner-product that is driven by the energy of input natural light and completed during the light field sensing process, it achieves real-time in-sensor computing. In this framework, the OCL has adaptive computing speed based on the imaging speed of the CIS. In other words, the faster the camera captures, the faster the OCL computes so that the OCL can always meet the computing requirements of real-world vision tasks. Moreover, the reduction in data throughput after the OCL is 96% so that the computational load of the electrical backend can be significantly reduced. On the other hand, incoherent natural light includes two spatial dimensions and one spectral dimension, the composition of matter can be identified and the mapping of its distribution in space can be realized by the SCNN, which starts a new paradigm for matter meta-imaging (MMI) beyond human eyes. To verify the capabilities of the proposed SCNN framework, we conducted several real-world complex vision tasks at video rate with the same SCNN chip, including pathological diagnoses with over 96% accuracy and anti-spoofing face recognition with almost 100% accuracy. Our implementation enables low-cost mass production and integration in the edge devices or cellphones of the proposed SCNN. Therefore, the proposed SCNN provides new MMI vision hardware and edge computing abilities for terminal artificial intelligence systems on diverse applications, such as intelligent robotics, industrial automation, medical diagnosis, and remote sensing.

Results

SCNN architecture

Our proposed SCNN consists of various spectral filters integrated on a CIS functioning as an on-chip analog OCL, followed by several electrical network layers (ENLs), as shown in Fig. 1b. Here, the spectral filters are designed to modulate light at different spectral and spatial points, which applies the convolutional kernel weights. Each spectral filter is completely aligned to a CIS pixel. $K = k \times k$ CIS pixels constitute a super-pixel and $N = n \times n$ super-pixels form an optical convolutional unit (OCU), as is shown in Fig. 1e. Furthermore, the entire OCL is an array of $H \times W$ OCUs. Because the OCUs are all identical, they perform spatial parallel analog 2D convolution calculations at different locations with megapixels.

Taking one OCU as an example (Fig. 1e), it has $K = k \times k$ convolutional kernels of size $n \times n$ and covers $n \times n$ super-pixels. The p-th $(1 \le p \le K)$ kernel has $N = n \times n$ weight vectors $w_{p1}(\lambda), w_{p2}(\lambda), \ldots, w_{pN}(\lambda)$, where $w_{pi}(\lambda) = t_{pi}(\lambda)r(\lambda)$ is determined by the transmission response $t_{pi}(\lambda)$ of the i-th filter in the kernel and the quantum efficiency $r(\lambda)$ of the CIS. Assuming that the input visual information represented by the superpixel is $x_i(\lambda)$ $(1 \le i \le N)$, the calculation result v_p of the kernel is as follows:

$$v_p = \sum_{i=1}^{N} I_{pi} = \sum_{i=1}^{N} \int_{\lambda_1}^{\lambda_2} x_i(\lambda) w_{pi}(\lambda) d\lambda = \sum_{i=1}^{N} \mathbf{w}_{ki}^T \mathbf{x}_i \tag{1}$$

where I_{pi} denotes the electrical signal output of the CIS pixel under the i-th filter in the kernel. Each OCU contains K kernels, and the OCL is a grid of $H \times W$ identical OCUs. Assume that the p-th kernel in the OCU located at $h, w(1 \le h \le H, 1 \le w \le W)$ has the output $v_{(h,w)p}$. Then, the 2D convolutional results of the OCL are:

$$\mathbf{F} = \{v_{(h,w)p}\} \in R^{H \times W \times K} \tag{2}$$

Therefore, OCL has K convolutional kernels of size $n \times n$ and stride $n \times n$. When $n = 1$, the OCL is a special convolutional layer with size 1×1 and stride 1×1, which can also be equivalent to a fully connected layer. When $n > 1$, the OCL is a strided convolutional layer with equal stride and kernel size, which can work as the combination of a convolutional layer and a pooling layer. Both the 1×1 convolutions and strided convolutions are widely adopted in CNNs such as ResNet[5]. Although the stride is restricted to be equal with kernel sizes, our experimental results have shown that our SCNN can still reach high performance for real-world tasks. In this way, the input visual signal has a spatial resolution of $nH \times nW$ and C spectral channel, which is equivalent to having $nH \times nW \times C$ voxels. C is determined by the sampling points in the spectral dimension. We assume that the light is locally homogeneous in one superpixel. The output feature map of the OCL has $H \times W$ spatial points and K channels. As usually $K \ll C$, the OCL can greatly compress the information in the spectral domain.

After in-sensor computing by the OCL, the output feature map is sent to the trained ENLs, which can comprise various ANN architectures such as FCNs and CNNs. Although the tailored OCL hardware is fixed after fabrication in our SCNN framework, its kernel size n and number of kernels $K = k^2$ can be reconfigured as well as $k \cdot n$ is fixed to the size of the OCU. A larger n leads to better capabilities of extracting spatial features and a larger k means more powerful spectral sensing abilities. Therefore, there is a trade-off between spatial and spectral features. We can choose the optimal value for k and n based on the actual needs of a specific task. Moreover, the ENLs can be changed and trained dynamically to suit different objectives. For example, in our disease diagnosis and face anti-spoofing tasks, we employed two different ENLs sharing the same OCL to perform pixel- and image-level predictions. Therefore, our SCNN framework combines the advantages of OCL by providing ultrafast sensing and processing of spatial and spectral features of natural images and the flexibility of ENLs with reconfigurable network designs for different tasks, enabling real-time MMI for different machine intelligent systems. Particularly, the OCL significantly reduces the computational load and data throughput of the electrical backends. The whole system can run in real time without the need for GPU. Therefore, the entire system is efficient and compact, which open the way for edge computing applications.

Metasurface based SCNN chip

In this work, we provide two implementations of the spectral filters for the SCNN. The first one is based on metasurfaces which provide flexible designed spectral modulation for the kernel weights of the OCL. Since different functions and applications require distinct metasurface designs to achieve the best results, we propose a gradient-based metasurface topology optimization (GMTO) algorithm to achieve an application-oriented metasurface design for different tasks such as thyroid disease diagnosis and anti-spoofing face recognition (Fig. 2a). Here, we first adopted freeform-shaped meta-atom metasurfaces[40] to generate millions of different

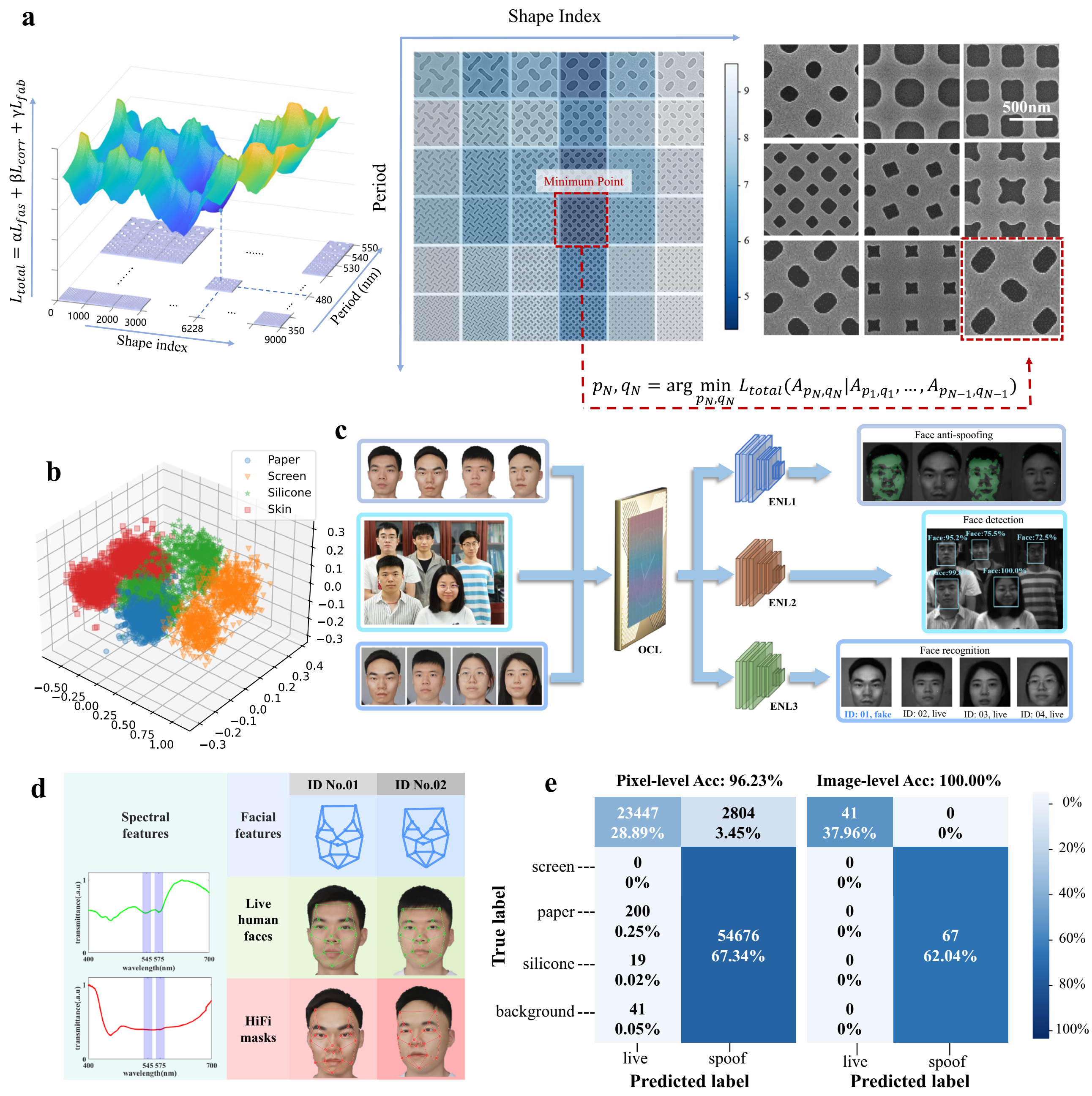
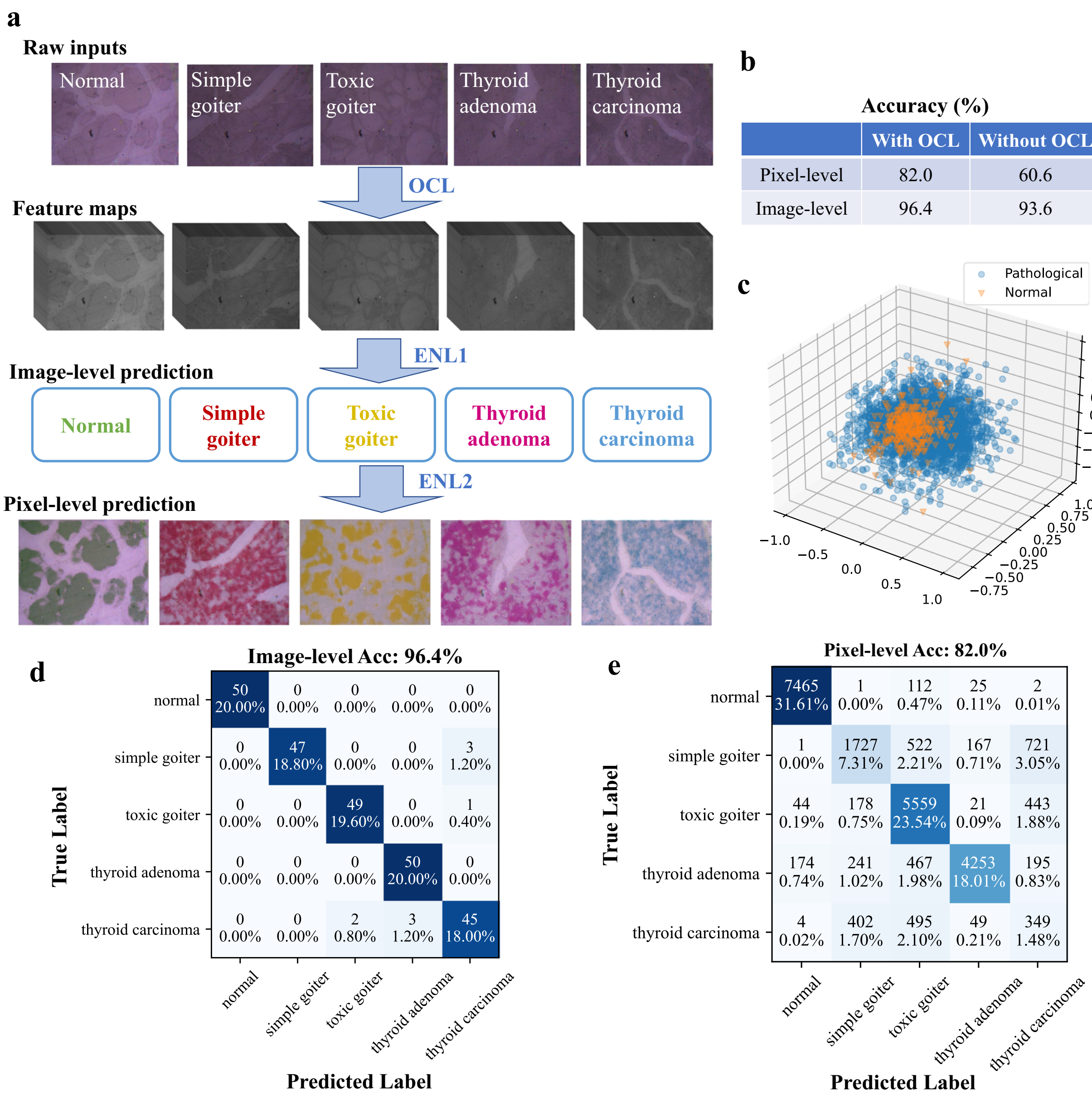


Fig. 2 | Metasurface based spectral convolutional neural network (SCNN) chip can be used for multiple vision tasks related to face recognition. a The gradient-based metasurface topology optimization (GMTO) algorithm is achieved by finding the minimum point of the designed loss function. **b** Spectral feature extraction results of the optical convolutional layer (OCL) visualized by PCA. Live skin and three spoof materials are separated. **c** The OCL has 9 kernels with size 1 × 1. By changing the electrical network layers (ENLs), the same SCNN chip can be trained to complete face anti-spoofing, face detection, and face recognition tasks. **d** Our SCNN chip can combine spectral features with spatial features and perform reliable anti-spoofing face recognition. **e** Confusion matrix for the pixel-level and image-level liveness detection results.

metasurface units and arranged all the metasurfaces into a 2D array. Thus, each metasurface unit can be uniquely represented by a pair of coordinates (p, q). To design N metasurfaces, the objective can be considered a function of $2N$ independent variables: $L(p_1, q_1, \ldots, p_N, q_N)$. We then utilized the GMTO algorithm to find the minimum points of $L(p_1, q_1, \ldots, p_N, q_N)$, obtaining the optimized design (see Supplementary Note 3 for details).

We found that OCL, designed by GMTO, could extract discriminating features with as few as nine kernels for live human skin and the thyroid tissue. The visualization results by principal component analysis (PCA)[41] are shown in Fig. 2b and Fig. 3c, respectively. Fewer kernels enable higher feature compression capability, higher spatial resolution, and lower computing costs for ENLs. Particularly, compared with our previous hyperspectral imaging works[34–37], SCNN uses very small number of metasurface units and provides an ONN-based approach for hyperspectral sensing, effectively avoiding the need for as many metasurface units as possible for high-precision spectral reconstruction (see Supplementary Note 1 for details). Finally, we implemented the OCUs with $H = 122$ and $W = 160$ by integrating millions of pixel-aligned metasurface units on top of a CIS (see "Methods" for details). The scanning electron microscopy (SEM) images of the fabricated metasurfaces are shown in Fig. 2a.

As is mentioned above, the size and number of convolutional kernels can be reconfigured. For example, the OCL shown in Fig. 1c can



Accuracy (%)

	With OCL	Without OCL
Pixel-level	82.0	60.6
Image-level	96.4	93.6

Image-level Acc: 96.4%

True Label \ Predicted Label	normal	simple goiter	toxic goiter	thyroid adenoma	thyroid carcinoma
normal	50 20.00%	0 0.00%	0 0.00%	0 0.00%	0 0.00%
simple goiter	0 0.00%	47 18.80%	0 0.00%	0 0.00%	3 1.20%
toxic goiter	0 0.00%	0 0.00%	49 19.60%	0 0.00%	1 0.40%
thyroid adenoma	0 0.00%	0 0.00%	0 0.00%	50 20.00%	0 0.00%
thyroid carcinoma	0 0.00%	0 0.00%	2 0.80%	3 1.20%	45 18.00%

Pixel-level Acc: 82.0%

True Label \ Predicted Label	normal	simple goiter	toxic goiter	thyroid adenoma	thyroid carcinoma
normal	7465 31.61%	1 0.00%	112 0.47%	25 0.11%	2 0.01%
simple goiter	1 0.00%	1727 7.31%	522 2.21%	167 0.71%	721 3.05%
toxic goiter	44 0.19%	178 0.75%	5559 23.54%	21 0.09%	443 1.88%
thyroid adenoma	174 0.74%	241 1.02%	467 1.98%	4253 18.01%	195 0.83%
thyroid carcinoma	4 0.02%	402 1.70%	495 2.10%	49 0.21%	349 1.48%

Fig. 3 | Experimental results of thyroid histological section diagnosis by the Metasurface based SCNN. a We exploit our SCNN to sense the raw datacube of thyroid histological section through a microscope. After the data are processed by the optical convolutional layer (OCL) and electrical network layers (ENLs), thyroid disease is automatically determined via image-level prediction. After the data are processed further by additional ENLs, the potential pathological areas are labeled in different colors via pixel-level prediction. **b** Without OCL, the classification accuracy based on the same monochromatic sensor decreases considerably for both image- and pixel-level predictions. **c** The spectral features from OCL can be visualized by principal component analysis (PCA). Normal and pathological tissues are separated. **d** Confusion matrix of the image-level thyroid pathology classification results of the SCNN chip on the test set. Our SCNN chip achieves 96.4% accuracy. **e** Confusion matrix of the pixel-level results. Our SCNN chip achieves 82.0% accuracy.

also be regarded as having 1 ($k=1, K=k^2=1$) convolutional kernel of size 3 × 3 and stride 3 × 3 ($n=3$). In this configuration, we need to sum the outputs of all of the CIS pixels in one OCU to generate an output feature map of size 160 × 122 × 1. We find this configuration performs worst in experiments because spectral features are more important than spatial features in the two applications. Therefore, we adopt the configuration of 9 convolutional kernels of size 1 × 1 and stride 1 × 1 to conduct further experiments.

To test the capabilities of the proposed SCNN framework, we employed the proposed SCNN for face anti-spoofing (FAS) to verify its performance. Nearly all of the current face recognition systems can be deceived by high-fidelity (HiFi) silicone masks, posing a great risk to privacy and security. However, when powered by our MMI, discriminative features can be extracted to detect HiFi masks. We captured images and obtained a test set containing 108 test samples from 31 different people, including several HiFi silicone masks, under natural light, and evaluated the performance of our SCNN chip. The results are shown in Fig. 2c, d. We can observe that our SCNN chip can effectively recognize live pixels, which are marked in green. Figure 2e shows the confusion matrix of the SCNN for all the test samples. The SCNN framework achieved 100% and 96.23% accuracy in image- and pixel-level liveness detection on our test dataset, demonstrating that our SCNN chip can achieve high reliability in anti-spoofing liveness detection applications (more results can be found in Supplementary Note 4). These results indicate the considerable potential for FAS systems.

Furthermore, we employed the designed SCNN chip to perform real-time anti-spoofing pixel-level liveness detection at different video frames. In this experiment, the entire system was run on a traditional Intel Core i5-6300HQ CPU, and the frame rate of the results was only limited by the CIS exposure time. The HiFi silicone masks can be easily detected at pixel level (more results can be found in the Supplementary Video 1). Thus, the proposed SCNN framework is expected to be widely used in the real-world applications of MMI. By simply redesigning and retraining the ENLs according to the needs of specific tasks, the function of the SCNN can be customized, such as face detection and recognition, as shown in Fig. 2c (more details of the redesigned ENLs can be found in Supplementary Note 5). The results show that the SCNN can accurately predict the locations of faces and achieve face recognition. This experiment indicates that the final output of the SCNN is highly customizable. The SCNN can flexibly adapt to various advanced CV tasks at video rates by simply changing and retraining the ENLs.

In addition to face anti-spoofing, we conducted automatic thyroid disease diagnosis experiments. The samples included normal thyroid tissue and tissues from four different diseases: simple goiter, toxic goiter, thyroid adenoma, and thyroid carcinoma. As shown in Fig. 3a, natural images of thyroid histological sections were first detected and processed using OCL. The feature maps output by the OCL is further processed by the ENLs to output the image-level thyroid disease classification results. Finally, the pixel-level disease detection results were output by other ENLs (see Supplementary Note 6 for details about the network). Figure 3d, e show that our SCNN framework can diagnose these four thyroid diseases, achieving an image-level testing accuracy of 96.4%, the ENLs only need 81.26 MOPs, more results can be found in Supplementary Note 9. Moreover, the SCNN chip automatically labeled the potential pathological areas in different colors at high spatial resolution. To study the role of the OCL, we conducted another experiment by replacing the OCL with a CIS without metasurfaces. After repeating the same ENLs training procedure, the image-level prediction accuracy decreased from 96.4% to 93.6%, and the pixel-level prediction accuracy decreased from 82.0% to 60.6% (Fig. 3b). The performance is much worse than using the OCL because OCL provides extra spectral sensing capabilities. Therefore, for the vision tasks related to spectral information, we need hyperspectral images rather than RGB images or grayscale images to get a good performance. If we complete the whole process by capturing data using a hyperspectral camera and implementing all neural network layers on the electrical computing platform, then we can get similar results compared with SCNN. However, the hyperspectral cameras usually have a very high cost and need time to scan a hyperspectral image. Moreover, the storing and processing cost of a hyperspectral image on an electrical computing platform is also very high (see Supplementary Note 2 for details). Therefore, conventional hyperspectral camera is not practical to be used in real-time edge computing applications, while the SCNN provides a simple but highly effective way to sense and process hyperspectral images for various portable terminals.

Pigments based SCNN chip with mass production

Besides metasurface-based spectral filters, we have also achieved the mass production of the SCNN on a 12-inch wafer utilizing pigments as spectral filters. The spectral filters are achieved by mixing several pigments with different organic solvents including ethyl acetate, cyclohexanone, and propylene glycol methyl ether acetate (PGMEA). The 12-inch wafer of the fabricated chips taped by lithography is shown in Fig. 4a. Each chip is only about $3 \times 3.5\ mm^2$ and can be integrated into any mobile device such as a smartphone to enable MMI. Focused ion beam-scanning electron microscope (FIB-SEM) image of the SCNN chip is shown in Fig. 4d. Each pigment-based filter is precisely aligned to a CIS pixel.

We selected 9 different pigments to form the spectral filters from several candidates that are compatible with lithography to make the differences between different targets in the feature maps outputted by the OCL as large as possible. Lithography enables large-scale integration of spectral filters, and the SCNN chip has a total of $400 \times 533 (H = 400, W = 533)$ superpixels. Therefore, the size of the feature map output by OCL is $400 \times 533 \times 9$. The spatial resolution is sufficient for most computer vision tasks, and OCL empowers massive parallel analog computing.

The fabricated chip is packaged into a tiny camera, as shown in Fig. 4b. The size of the camera is approximately $6.5 \times 7mm^2$. We placed the pathological thyroid sections immediately above the camera lens without any microscope, which is impossible for traditional pathological diagnosis. Natural images of thyroid histological sections were first obtained and processed using OCL. The feature maps output by the OCL is then further processed by the ENLs to output the image-level thyroid disease classification results. The camera can capture only a blurry image rather than a sharp image showing clear textures since a microscope is not used. Some samples of pathological sections and their feature maps outputted by the OCL are shown in Fig. 4f. The feature maps display few spatial features. However, we still reach a classification accuracy of 96.46%. Furthermore, we also conducted another experiment by replacing OCL with CIS without pigment-based filters to study the role of OCL. After repeating the same data collection and ENL training procedure, the classification accuracy decreased from 96.46% to 47.09%. The tiny size of the finished camera allows it to be integrated into various medical instruments such as laparoscopes. Thus, the proposed SCNN framework shows considerable potential as an ancillary diagnostic tool in clinical medicine and might assist doctors in precisely localizing lesions in real-time during surgery.

We have also achieved the face anti-spoofing task using the pigment-based SCNN (Fig. 4g). The confusion matrix of the classification results and more experimental results can be found in Supplementary Note 7 and Supplementary Video 2. Compared with metasurface-based SCNN, pigment-based SCNN achieved mass production by lithography, thus obtaining high integration and high spatial resolution. However, the metasurfaces can provide more powerful light field modulation capabilities and greater design freedom[36–39,42], resulting in more spectral information and more space for customization. Based on the concept of the SCNN, the metasurface-based architecture also has further potential in sensing and processing other light dimensions, e.g., polarization and phase[43–46]. Besides, metasurfaces also have the potential to achieve mass production via standard semiconductor lithography process. Therefore, in practical, we can choose and design the optimal SCNN chip depending on the specific requirements of the application. It can be predicted that SCNN chips will have more potential in various applications.

Discussion

We proposed an integrated SCNN framework that achieves in-sensor edge computing of incoherent natural light. It can detect visual information in natural raw 3D datacube with both spatial and spectral features by performing optical analog computing in real-time. Leveraging both the OCL and ENLs, SCNN can achieve high performance even on edge devices with limited computing capabilities, which enables edge computing with MMI functions. In practical applications, utilizing the high versatility of ENLs, a specific SCNN chip can be easily adapted to various advanced vision tasks as demonstrated in this work. For the OCL, it is designed to perform inferencing for spectral sensing and computing in edge devices rather than in-situ training. Therefore, for a specific application, the weights can be fixed. To achieve a completely new task at high performance, we need to re-design and re-fabricate the chip. For optical neural networks (ONNs) with weights encoded by non-tunable optical structures, we can adopt a similar

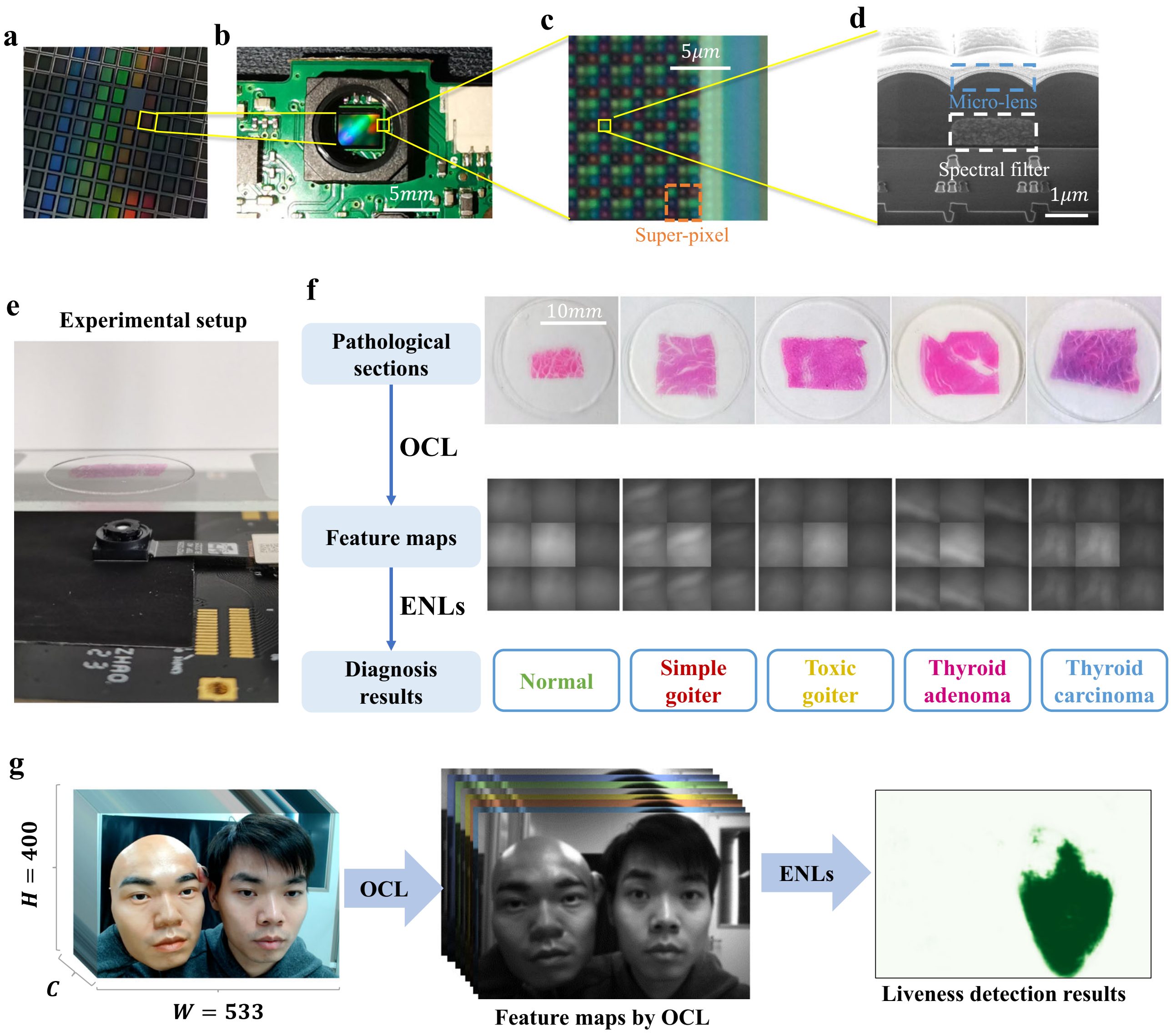


Fig. 4 | Spectral convolutional neural network (SCNN) chip implemented by utilizing pigments as spectral filters and achieving mass production on a 12-inch wafer. a The fabricated SCNN chips on a 12-inch wafer by lithography. **b** A tiny camera equipped with the SCNN chip. It can achieve in-sensor edge computing and spectral sensing. The size of the SCNN chip is only about $3 \times 3.5mm^2$ and the size of the whole camera is about $6.5 \times 7mm^2$ **c** A microscope image of the fabricated SCNN chip. It has 9 convolutional kernels of size 1×1 and stride 1×1. A super-pixel contains 9 image sensor pixels. **d** The focused ion beam-scanning electron microscope (FIB-SEM) image of the SCNN chip. One image sensor pixel is covered by a pigment-based spectral filter and a micro-lens. The fabrication process is completely standard semiconductor lithography process. **e** We place the thyroid pathological sections right above the lens without a microscope. **f** The sections and the corresponding feature maps outputted by optical convolutional layer (OCL). **g** The face anti-spoofing results of the pigment-based SCNN. ENL: Electrical convolutional layer.

strategy as refs. [21,22,24,29], which is to design the network by electrical computing and then fabricate the optical computing layer for specific tasks in terminal devices for edge computing. It is a tailored chip for a specific task for edge computing applications. The computing speed and power consumption of OCL depend only on the exposure time and the power of the CIS, empowering ultrafast optical computing at high energy efficiency.

To achieve hyperspectral imaging and sensing, we can also adopt a conventional hyperspectral camera to scan hyperspectral images, and then process the images on GPU. However, such a system cannot be integrated on edge devices because GPU has large size, high energy consumption, and high cost that cannot meet the requirements of edge devices with limited computing capabilities. Besides, the conventional hyperspectral camera is also bulky, expensive, and not capable of real-time imaging. Our OCL is in-sensor computing that provides a substantial reduction of 96% in data throughput. The computing speed of OCL only depends on the imaging speed of the CIS. The faster the CIS captures, the faster the computing speed of OCL can be. Therefore, the OCL can always satisfy the computing requirements of real-world tasks. Besides, the SCNN makes it possible to process hyperspectral images using only a few extra digital neural network layers on edge devices. It can empower edge devices with both sensing and computing capabilities for various real-world complex vision tasks.

Compared with existing on-chip works, as is shown in Fig. 1 and Table 1, our SCNN can process natural hyperspectral images with high spatial resolution/pixels. It does not rely on coherent light sources, fiber coupling, or waveguide delay. Although CIS is relatively slow

Table 1 | Comparison with existing on-chip ONN works

Publication	Pixels	Computing speed	Computing density	In-sensor	Incoherent light	MMI	Application
X., X. et al.[27] Nature, 2021	500 × 500	1.785 TOPS	–	×	×	×	handwritten digit recognition (HDR)/image processing
F., J. et al.[26] Nature, 2021	128 × 128	4 TOPS	1.2 TOPS/mm2	×	×	×	HDR/edge detection
A., F. et al.[11] Nature, 2022	5 × 6	0.27 TOPS	3.5 TOPS/mm2	×	×	×	low-resolution image classification
F., T. et al.[24] Nat. Commun., 2023	28 × 28	13.8 POPS	–	×	×	×	HDR
M., X. et al.[31] Nat. Commun., 2023	28 × 28	0.27 TOPS	25.48 TOPS/mm2	×	×	×	HDR
B., B. et al.[32] Nat. Commun., 2023	250 × 250	–	1.04 TOPS/mm2	×	×	×	HDR/edge detection
D., B. et al.[33] Nature, 2024	28 × 28	0.108 TOPS	–	×	×	×	HDR
Ours	400 × 533	Adaptive[a]	Adaptive	√	√	√	complex tasks in the real world: face anti-spoofing and disease diagnosis

[a] OCL performs analog computing. It has adaptive computing speed that can always meet the imaging speed.

compared with the commonly used high-speed photodetector, we still achieve considerable computing speed and density compared with existing photodetector-based works because CIS has high integration and can take full advantage of space division multiplexing. If we replace the CIS with high-speed PD array, there is still great potential for improvement in computing speed. More detailed analysis about computing speed can be found in Supplementary Note 2. Actually, as CIS is the most integrated optoelectronic device, we can have hundreds of millions of pixels at a very low cost. The SCNN provides the strategy of utilizing every single pixel to perform optical computing via CIS to achieve high computing density and reduce the number of photoelectronic conversions. Based on the above advantages of SCNN architecture, we have achieved mass production on a 12-inch wafer of the pigment-based SCNN. Thus, the proposed SCNN opens a new practical in-sensor computing platform for complex vision tasks with MMI functions in the real world.

Methods

Fabrication of the metasurface-based SCNN Chip

The designed metasurfaces were formed using EBL on a silicon-on-insulator (SOI) chip. The silicon layer was 220 nm thick. The metasurface patterns were transferred onto the silicon layer via inductively coupled plasma etching (ICP). To remove the silicon layer from the underlayer, buffered hydrofluoric acid was used to wet etch the silicon dioxide layer. Finally, the entire top Si layer with the designed metasurfaces was transferred and attached to the surface of the CIS using polydimethylsiloxane (PDMS). We used a Thorlabs CS235MU camera for CIS. The proposed SCNN chip can be fabricated using a CMOS-compatible process and can be mass-produced at low cost.

Fabrication of the pigment-based SCNN chip

The pigment-based SCNN Chip is produced at semiconductor foundry on a 12-inch wafer, employing a standard color filter array process via I-line lithography. The CIS wafer is uniformly coated with a color resist. To render the pattern insoluble, it is UV-cured by exposure through a carefully designed photomask. Subsequent to this, any unnecessary portions of the color resist are removed using the developing solution. Following this removal, the pattern is further solidified through a baking process. This comprehensive sequence of steps is repeated nine times.

Following the color filter layer process, a planarization layer was established using the Chemical Mechanical Polishing (CMP) technique to ensure a flat and uniform surface. Subsequently, a photoresist layer was uniformly applied onto this planarized surface using a spin-coating method. This photoresist layer was patterned by UV light exposure through a predefined mask. The excess photoresist was then removed in a development process, leaving behind the desired patterns. The wafer was subjected to a reflow baking process, during which the patterned photoresist naturally reflowed into the shape of microlenses, driven by surface tension and thermodynamic effects.

Implementation of the ENLs

The ENLs in the SCNN are realized using the TensorFlow[47] framework and trained on an NVIDIA RTX3080 GPU. Several volunteers have participated in the face anti-spoofing task. The authors affirm that human research participants provided informed consent for publication of the images in Figs. 2, 4. The study is conducted under the guidelines provided by Tsinghua Ethics Committee. Additional implementation and training details of ENLs are provided in Supplementary Note 6 and 7. After training, the ENLs and OCL formed a fully functional SCNN. The electrical components of the SCNN were run on an Intel Core i7-11700 @2.5 GHz CPU for real-time applications.

Data availability

The data for Figs. 1c, 2b, 3c are provided in Source Data file. The authors declare the other data supporting the findings of this study are available within the paper (and its supplementary information files). Source data are provided with this paper.

Code availability

We have developed codes for training the ENLs. A surrogate forward prediction model is also designed to fast predict the transmission responses of meta-atoms. The codes and detailed information can be found at our GitHub Page[48] (https://github.com/rao1140427950/scnn_mpcf). Other algorithms and methods are included in this published article (and its supplementary information files).

References

1. OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
2. Cao, Y. et al. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226* (2023).
3. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
4. Dietler, N. et al. A convolutional neural network segments yeast microscopy images with high accuracy. *Nat. Commun.* **11**, 5723 (2020).

5. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* (Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016).
6. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional networks for biomedical image segmentation* (Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015).
7. Girshick, R. *Fast r-cnn* (Proceedings of the IEEE International Conference on Computer Vision, 2015).
8. Sui, X., Wu, Q., Liu, J., Chen, Q. & Gu, G. A review of optical neural networks. *IEEE Access* **8**, 70773–70783 (2020).
9. Miller, D. A. Are optical transistors the logical next step? *Nat. Photonics* **4**, 3–5 (2010).
10. Caulfield, H. J. & Dolev, S. Why future supercomputing requires optics. *Nat. Photonics* **4**, 261–263 (2010).
11. Ashtiani, F., Geers, A. J. & Aflatouni, F. An on-chip photonic deep neural network for image classification. *Nature* **606**, 501–506 (2022).
12. Huang, C. et al. Prospects and applications of photonic neural networks. *Adv. Phys. X* **7**, 1981155 (2022).
13. Shi, W. et al. LOEN: Lensless opto-electronic neural network empowered machine vision. *Light Sci. Appl* **11**, 121 (2022).
14. Wang, T. et al. An optical neural network using less than 1 photon per multiplication. *Nat. Commun.* **13**, 123 (2022).
15. Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58 (1994).
16. Shen, Yichen et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
17. Zhang, H. et al. An optical neural chip for implementing complex-valued neural network. *Nat. Commun.* **12**, 457 (2021).
18. Ribeiro, A. et al. Demonstration of a 4×4-port universal linear circuit. *Optica* **3**, 1348–1357 (2016).
19. Xu, Z. et al. Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence. *Science* **384**, 202–209 (2024).
20. Chen, Y. et al. All-analog photoelectronic chip for high-speed vision tasks. *Nature* **623**, 48–57 (2023).
21. Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
22. Luo, Y. et al. Design of task-specific optical systems using broadband diffractive neural networks. *Light Sci. Appl* **8**, 112 (2019).
23. Qian, C. et al. Performing optical logic operations by a diffractive neural network. *Light Sci. Appl* **9**, 59 (2020).
24. Fu, T. et al. Photonic machine learning with on-chip diffractive optics. *Nat. Commun.* **14**, 70 (2023).
25. Goi, E., Schoenhardt, S. & Gu, M. Direct retrieval of Zernike-based pupil functions using integrated diffractive deep neural networks. *Nat. Commun.* **13**, 7531 (2022).
26. Feldmann, J. et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **589**, 52–58 (2021).
27. Xu, X. et al. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **589**, 44–51 (2021).
28. Wang, T. et al. Image sensing with multilayer nonlinear optical neural networks. *Nat. Photon.* **17**, 408–415 (2023).
29. Zheng, H. et al. Meta-optic accelerators for object classifiers. *Sci. Adv.* **8**, eabo6410 (2022).
30. Wu, H. & Dai, Q. Artificial intelligence accelerated by light. *Nature* **589**, 25–26 (2021).
31. Meng, X. et al. Compact optical convolution processing unit based on multimode interference. *Nat. Commun.* **14**, 3000 (2023).
32. Bai, B. et al. Microcomb-based integrated photonic processing unit. *Nat. Commun.* **14**, 66 (2023).
33. Dong, B. et al. Partial coherence enhances parallelized photonic computing. *Nature* **632**, 55–62 (2024).
34. Wu, C. et al. Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network. *Nat. Commun.* **12**, 96 (2021).
35. Xu, S. et al. High-order tensor flow processing using integrated photonic circuits. *Nat. Commun.* **13**, 7970 (2022).
36. Pi, L. et al. Broadband convolutional processing using band-alignment-tunable heterostructures. *Nat. Electron* **5**, 248–254 (2022).
37. Xiong, J. et al. Dynamic brain spectrum acquired by a real-time ultraspectral imaging chip with reconfigurable metasurfaces. *Optica* **9**, 461–468 (2022).
38. Rao, S., Huang, Y., Cui, K. & Li, Y. Anti-spoofing face recognition using a metasurface-based snapshot hyperspectral image sensor. *Optica* **9**, 1253–1259 (2022).
39. Yang, J. et al. Ultraspectral imaging based on metasurfaces with freeform shaped meta-atoms. *Laser Photonics Rev.* **16**, 2100663 (2022).
40. Yang, J. et al. Deep-learning based on-chip rapid spectral imaging with high spatial resolution. *Chip* **2**, 100045 (2023).
41. Abdi, H. & Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 433–459 (2010).
42. Ding, F., Meng, C. & Bozhevolnyi, S. I. Electrically tunable optical metasurfaces. *Photonics Insights* **3**, R07–R07 (2024).
43. Ji, J. et al. On-chip multifunctional metasurfaces with full-parametric multiplexed Jones matrix. *Nat. Commun.* **15**, 8271 (2024).
44. Shen, Z. et al. Monocular metasurface camera for passive single-shot 4D imaging. *Nat. Commun.* **14**, 1035 (2023).
45. Zaidi, A. et al. Metasurface-enabled single-shot and complete Mueller matrix imaging. *Nat. Photon.* **18**, 704–712 (2024).
46. Li, T. et al. Revolutionary meta-imaging: from superlens to metalens. *Photonics Insights* **2**, R01–R01 (2024).
47. Abadi, M. et al. *Tensorflow: a system for large-scale machine learning*. (Osdi Ser. 16, Savannah, GA, USA, 2016).
48. Cui, K. et al. *Spectral Convolutional Neural Network Chip for In-sensor Edge Computing of Incoherent Natural Light*. rao1140427950/scnn_mpcf, https://doi.org/10.5281/zenodo.14258334 (2024).

Acknowledgements

This work is supported by the National Key Research and Development Program of China: 2023YFB2806703 (Y. H.), 2022YFF1501600 (K. C.). The National Natural Science Foundation of China: Grant No. U22A6004 (Y. H.); Beijing Frontier Science Center for Quantum Information; and Beijing Academy of Quantum Information Sciences. The authors would like to thank Prof. Chengwei Qiu for his valuable discussion on the manuscript. And thank Dr. Jiawei Yang, Dr. Jian Xiong, and Chenxuan Wang for their help and advice on this study.

Author contributions

K. C., S. R. contributed equally to this work. K. C. conceived the study and proposed the optical computing strategy. S. R. completed the SCNN framework and conducted the experiments with the help of S. X. Y. H. supervised the project and advised on device optimization. X. C., Z. H., and Y. W. completed the chip fabrication. S. W. and Y. L. advised on network design and training. S. R. and K. C. wrote the manuscript with contributions from all other coauthors. F. L., X. F., and W. Z. provided useful commentary on the results. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-55558-3.

Correspondence and requests for materials should be addressed to Kaiyu Cui or Yidong Huang.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints