nature portfolio

Peer Review File

Spectral Convolutional Neural Network Chip for In-sensor Edge Computing of Incoherent Natural Light

Corresponding Author: Professor Kaiyu Cui

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have presented an approach to perform optical preprocessing of natural images in preparation for further electronic neural network processing. Their approach is able to make use of additional spectral information typically lost on conventional color camera sensors through custom manufactured color filters aligned to a CMOS sensor. They have demonstrated two real-world applications in the form of pathology slide classification and face anti-spoofing. The concept is interesting with potential for applications in real-time edge computing, however I have several major reservations about the implementation presented and claims made in the manuscript.

Major comments:
1. The authors claim to implement optical convolution layers but their implementation uses convolutional kernels of size 1x1 and stride 1x1. This is the special degenerate case and the authors have not demonstrated spatial convolution using their approach. It is not reasonable to make claims about analog 2D convolutions.

2. The experimental setup as presented has major overlap with existing multispectral filter arrays (e.g. Lapray et. al. Sensors 2014) and hyper-spectral imaging in general.

3. The computational benefits are not discussed technically. What fraction of the computation does the optical layer perform as opposed to the electronic neural layers? What is the total computational throughput of the optical layers and how does this compare to other approaches? What is the total reduction in data throughput?

Minor comments:
1. The authors claim existing optical neural networks only work on coherent input light, however Wang et al 2023 makes use of incoherent natural input light.
2. How would the optical convolutional layer architecture be setup to perform spatial convolution?
3. How do the results in Figure 3 compare to a conventional camera with 3 color channels?
4. Supp. Figure 1 labels says the convolution is 3x3 but the text claims 1x1.
5. Could you elaborate more about the datasets used for the experiments?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Summary: This paper proposes a free-space optical dot-product engine for analog image pre-processing, combined with later digital CNN; it can realize high accuracy in 2 selected classification tasks. Experimental demonstration has shown its usage in image classification tasks.

Comments
1. The novelty and intellectual contribution are limited from the circuit/architecture/algorithm design sides. The engineering efforts in building such a demo are appreciated. Extensive designs and experimental demonstrations have been conducted on free-space optical dot-product engines, convolution engines, diffractive neural networks, etc. Some of them are even multi-layer, nonlinear, reconfigurable, and working with visible light, and can handle phase/polarization. A thorough comparison to prior free-space optics and integrated photonic accelerator designs is needed, especially on cost, efficiency, speed, throughput, size, reconfigurability, reliability, robustness, expressivity, etc.
2. The demonstrated optical convolution unit is equivalent to a Conv2d(in_channels=1, out_channels=9, kernel_size=1, stride=1, bias=False) or Conv2d(in_channels=1, out_channels=1, kernel_size=3, stride=3, bias=False) with fixed, quantized, noisy, positive-value weights. The expressivity of such a CONV layer is a concern, given that advanced real-world CV tasks require much more complicated DNN models. By replacing the first CONV of a DNN with this optical CONV, what are the overall impacts and system-level benefits?
3. The reconfigurability is big concern, how to enable reconfigurable weights? And what are the underlying trade-offs?
4. The claimed reconfigurable kernel size sacrifices the efficiency by electronically summing the dot-product results, which might not be very efficient.
5. What is the robustness of such a system? Quantization, resolution, fabrication variation, thermal sensitivity, alignment sensitivity, signal-to-noise ratio, etc.
6. The output feature map is 400x533x9, which is a large feature map. What is the data movement cost, bandwidth requirement, and system throughput?
7. How does that compare to standard CNN taking the RGB image as inputs?
8. If the main advantage is from the MPCF in processing multiple spectrums of the image, how does it compare to metasurface-based DONN that can sense and process other dimensions, e.g., polarization and phase?
9. By checking the code provided, the images are preprocessed and stored as 9-channel input features, and pass through a very deep CNN and ResNet, which makes the initial optical CONV almost meaningless. Why not just input the raw features from the sensor to the used large digital full-precision CNN running on GPU? The claim that it is very efficient on edge devices without the need of GPU is not very justified.

(Remarks on code availability)
The code is a standard Keras-based CNN example for image classification.

Reviewer #4

(Remarks to the Author)
The authors in this work proposed an integrated spectral convolutional neural network (SCNN) framework with in-sensor computing capability to detect visual information in broadband natural incoherent light. Thus, the computing speed can be improved and the energy efficiency is enhanced. The results are interesting. The authors are suggested to address my concerns before the manuscript being published.

(1) What is the computing speed and power consumption of the proposed SCNN chip, is it ahead of existing architectures?
(2) The authors noted that CNNs require significant computational resources. Is SCNN more lightweight? Please provide a quantitative analysis.
(3) In Figure 3, how is Acc calculated? Are there other more diverse evaluation metrics available? If so, please provide an analysis.
(4) In the tasks of Histological Section Diagnosis and Face Anti-spoofing, how do existing methods perform? What are the advantages of the method proposed by the authors compared to existing methods? Please provide a detailed explanation.
(5) In this manuscript, a neural network chip with photoelectric hybrid architecture is proposed, a comparison table including key parameters with existing on-chip neural network should be provided.
(6) If available, please provide both quantitative and qualitative comparisons with existing methods.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you to the authors for responding to the review comments and concerns. Thank you for the effort made to address all the comments; however, the authors' rebuttal does not adequately resolve two remaining points. The argument in the revised manuscript has the following problems, especially 1 questions the validity of the claims:

1) The calculation of the number of operations performed is suspect. First, the authors should first demonstrate that the described compression rate of 4.16% results in no loss performance for the implemented tasks. Second, the throughput of 2NC/T does not use an appropriate sampling time. The camera sensor as detailed is the CS235MU with a maximum full frame-rate of 165.5fps. From this readout, it is inappropriate to use T=0.027ms, and instead T = 6ms is more appropriate. The comparison to a high-speed PD array for a throughput of 107 TOPS is also inappropriate as the corresponding data rate would be 40000TB/S, and that doesn't even include a host of other problems with light intensity and electronics constraints.

2) In the text, the kernel sizes are written: "its kernel size n and number of kernels $K = k2$ can be reconfigured as well as $k \cdot n$ is fixed to the size of the OCU," along with "Therefore, OCL has KK convolutional kernels of size $n \times n$ and stride $n \times n$." While this is an accurate description of the system, it obfuscates the point that the sensor size is fixed and the input light is filtered by the SCNN without the possibility for additional spatial mixing, as would be expected for a convolutional layer. As the primary successful implementation of the SCNN in the manuscript was using nine designed spectral filters, I believe an emphasis of the SCNN as primarily a high-speed customizable hyper-spectral imaging method would only serve to strengthen the manuscript.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
1. For each hyperspectral image, the number of operations saved is 569.3M operations, which is only the operations for a standard 3x3 Conv2d in a DNN. Compared to the rest of the network layers, it is not a significant computation reduction.
2. The claimed 10^7 TOPS using PD arrays is not valid, as no data movement solutions can support such a data readout rate.
3. The reconfigurability concerns of this fixed processor are not addressed. A technological solution to enable reconfiguration is required. It has nothing to do with training, just for other functionality to use this device. Otherwise, fixed functionality" should be put in the title.
4. The authors claimed the weights are not quantized and noisy, which is not true. The fixed weights are from fabrication, it has to have certain precision and process variation.

(Remarks on code availability)
The codes do not contain much photonic analog part. It is a pure digital CNN training code.

Reviewer #4

(Remarks to the Author)
This manuscript has been completely revised based on the recommendations made. I recommend accepting it.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you to the authors for the revisions in response to the previous round of reviews. One concern is that the paper makes

claims that overstate what has actually been shown and the text needs to be toned down. This also seems to have been reflected in another review report. The authors have made changes to remedy some of these statements, but further changes are needed: Specifically, it is necessary for the authors to display the actual demonstrated performance of the experimental device (17.2 GOPS) and corresponding compute density in Table 1. The current number displays a potential performance (21.0 TOPS) that cannot be realized with the hardware used in the manuscript. This is misleading. Please update Table 1, as described above, to reflect what you actually show in the paper.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
Thanks for the response from the authors.
I agree with Review 1 that the proposed SCNN is a customized (in the sense of fixed function after fab) spectral imaging/preprocessing method to collect information from multiple spectrums. It is not questionable and is better than collecting only RGB channels. However, selling this chip (3x3/1x1 conv) as an edge NN accelerator that speeds up the whole NN system will have a lot of problems in speed, data movement, reconfigurability, etc.
No further follow-on questions.

(Remarks on code availability)

Responses to Reviewer #1:

Comments: *The authors have presented an approach to perform optical preprocessing of natural images in preparation for further electronic neural network processing. Their approach is able to make use of additional spectral information typically lost on conventional color camera sensors through custom manufactured color filters aligned to a CMOS sensor. They have demonstrated two real-world applications in the form of pathology slide classification and face anti-spoofing. The concept is interesting with potential for applications in real-time edge computing, however I have several major reservations about the implementation presented and claims made in the manuscript.*

Response: We appreciate the referee's thorough review, accurate summary, and valuable concerns on our work. Your comments are very helpful for our improvement, and we have already added additional theoretical analysis and experimental results to demonstrate the implementation and claims of this work. Below, we address each of the raised concerns in detail.

Major comment 1: *The authors claim to implement optical convolution layers but their implementation uses convolutional kernels of size 1x1 and stride 1x1. This is the special degenerate case and the authors have not demonstrated spatial convolution using their approach. It is not reasonable to make claims about analog 2D convolutions.*

Response: We appreciate the referee's comments. We have illustrated in the manuscript (page 7 lines 153-158) that "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 optical convolutional unit (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." The kernel size is not limited to be 1×1. In our implementation, we designed 9 different spatial filters. As is claimed in the manuscript (page 8 lines 195-202) the OCL can be regarded to have 9 kernels of size 1×1 or 1 kernel of size 3×3. It depends on how we sum the pixel values of the image sensor.

As is described on page 6 lines 128-132 in the Results/SCNN Architecture Section, 2D convolution requires sliding the kernel to the different spatial locations. In the OCL, the kernels are realized by spectral filters that cannot be moved. Therefore, we put the same kernel at different spatial locations (that is, the 2D OCU array) to replace the sliding operation. In this way, the same kernel can also be applied to different spatial locations. Therefore, it can fully take advantage of space division multiplexing and achieve large-scale integration, which leads to the high spatial resolution (also discussed on page 16 line 375). However, the stride of the kernel needs to be equal to the size of the kernel. In all, we can implement the OCL with $K = k^2$ kernels, $n \times n$ size and $n \times n$ stride.

Major comment 2: *The experimental setup as presented has major overlap with existing multispectral filter arrays (e.g. Lapray et. al. Sensors 2014) and hyper-spectral imaging in general.*

Response: We appreciate the referee's comments. It's our negligence that we didn't state the difference between our work and hyperspectral imaging clearly. The detailed comparison between

SCNN and hyperspectral imaging is added on page 3 lines 80-91 of the revised manuscript and Supplementary Note 1. Although both this work and previous spectral imaging works utilize spectral filters and image sensors, they have different architecture and designing concept to achieve different capabilities. Previous hyperspectral imaging works adopted spectral filters as the sensing matrix and got the compressively sensed hyperspectral images. "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."

From the perspective of practical capabilities, we can compare this work with our previous work on the hyperspectral imaging for face anti-spoofing task (Shijie et al, Optica, 2022). Empowered by the new SCNN framework, this work is completely beyond the previous work at the spatial resolution, system simplicity, running efficiency, and anti-spoofing performance The comparison table between hyperspectral imaging and SCNN is listed below. Detailed analysis has been added in Supplementary Note 1.

	Hyperspectral imaging	**SCNN**
Number of different metasurfaces	49~400	**9**
Spatial resolution (pixels)	5×10^4	$\mathbf{2.13 \times 10^5}$
Frames per second (fps)	<0.1	**>10**
FAS accuracy (%)	~95%	**~99%**
Spectrum reconstruction	√	×
Pixel-level real-time sensing	×	√
Edge computing	×	√

Major comment 3: *The computational benefits are not discussed technically. What fraction of the computation does the optical layer perform as opposed to the electronic neural layers? What is the total computational throughput of the optical layers and how does this compare to other approaches? What is the total reduction in data throughput?*

Response: Thanks for your suggestions and questions. We have realized the lack of quantitative analysis and comparisons with other works. The quantitative analysis of the computing speed and fully comparation with existing works are added in Supplementary Note 2 and Table 1 in the revised manuscript of pages 16-17, as is shown below.

If we replace the OCL with a digital convolutional layer, then the digital layer has to process the multi-channel and high-resolution hyperspectral images, which is computationally expensive and brings great difficulties to edge computing. The OCL can reduce about 569.3M operations for processing. It is in-sensor computing that provides a computing speed as high as 21.0 TOPS, so that the computational load of the electrical backend can be significantly reduced (see detail in Supplementary Note 2 and Table 1). Moreover, the exposure time of the CIS is relatively low

(sampling rate is about 37 kHz) compared with high-speed photodetector (sampling rate can exceed 100 GHz). If we replace the CIS with PD array, the computing speed can be further improved to over 10^7 TOPS. And the OCL can reduce hyperspectral images into 9-channel feature maps, which is a 96% reduction in data throughput required for transferring hyperspectral images. In addition, to acquire such hyperspectral images, an extra hyperspectral camera is needed and high-performance electronic computing platforms, such as graphic process units (GPU) are inevitable. Therefore, the quantitative analysis indicates that our OCL can significantly reduce the computational burden and data throughput for processing hyperspectral images, thus realize the in-sensor edge computing abilities which is impossible for previous hyperspectral imaging task.

Compared with other optical computing approaches, our OCL does not rely on coherent light sources, fiber coupling, or waveguide delay, but it provides the sensing and processing capabilities of hyperspectral images at high spatial resolution. We can adopt the proposed OCL on complex real-world tasks far beyond handwritten digit recognition.

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 digits 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**	**21.0 TOPS**	**5.3 TOPS/mm2**	√	√	√	**complex tasks in the real world: face anti-spoofing and disease diagnosis**

MMI: Matter Meta-Imaging

Minor comment 1: *The authors claim existing optical neural networks only work on coherent input*

light, however Wang et al 2023 makes use of incoherent natural input light.

Response: We appreciate your concern regarding the work on coherent input light. The comment guided us to revise the description in the manuscript more rigorously: "existing on-chip OCNNs hardly accept broadband incoherent natural light" (page 2 line 58 of the revised manuscript). In fact, we are the first to achieve integrated optical computing utilizing natural light to the best of our knowledge. The comparison with these integrated OCNN works is also shown in Table 1 of the revised manuscript. . Although some previous optical neural networks such as Wang et al 2023 do not rely on coherent light, their optical computings are performed on spatial light rather than integrated architectures. Moreover, these works regard the incident light as monochrome images and the spatial resolution is limited. That is, they do not have the capabilities of sensing incoherent natural light.

Minor comment 2: *How would the optical convolutional layer architecture be setup to perform spatial convolution?*

Response: Thanks for the question. As mentioned in Response 1, the spatial convolution is achieved by the 2D OCU array, which is equivalent to moving the convolution kernel to different spatial locations. Therefore, it can fully take advantage of space division multiplexing and achieve large-scale integration, which leads to the high spatial resolution (also discussed on page 16 line 375)

Minor comment 3: *How do the results in Figure 3 compare to a conventional camera with 3 color channels?*

Response: We have added related descriptions in our revised manuscript (page 12 lines 272-274) RGB sensor cannot provide spectral sensing capabilities. In the RGB color space, the live and spoof faces (or sections of tissues with different diseases) are not distinguishable. Therefore, common RGB sensors cannot achieve such tasks.

Minor comment 4: *Supp. Figure 1 labels says the convolution is 3x3 but the text claims 1x1.*

Response: We apologize for the confusion caused by the mistakes in the figures. The Suppl. Figure 1 is mistaken. We have corrected it.

Minor comment5: *Could you elaborate more about the datasets used for the experiments?*

Response: Thank you for the detailed suggestion. We have added these information to Supplementary Note 7. "For face anti-spoofing, we collected more than 200 samples of live subjects (from 5 different people) and spoof subjects (from 15 different masks). Then the dataset was split into training set and testing set at a ratio of 4:1 according to the subject identities. For disease diagnosis, we collected about 500 samples from 100 different sections. 80 sections were employed as training set and the remaining 20 sections were testing set."

To reviewer #2:

Comments: *I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.*

Response: Thanks for the time and patience in reviewing our manuscript. We have made major revisions to our manuscript and supplementary material, which includes:

1. Add new experimental results about metasurface-based OCL. We present the design, fabrication, and performance of the metasurface-based OCL (Fig. 2 and Fig. 3) to further clarify the advances of the proposed SCNN:

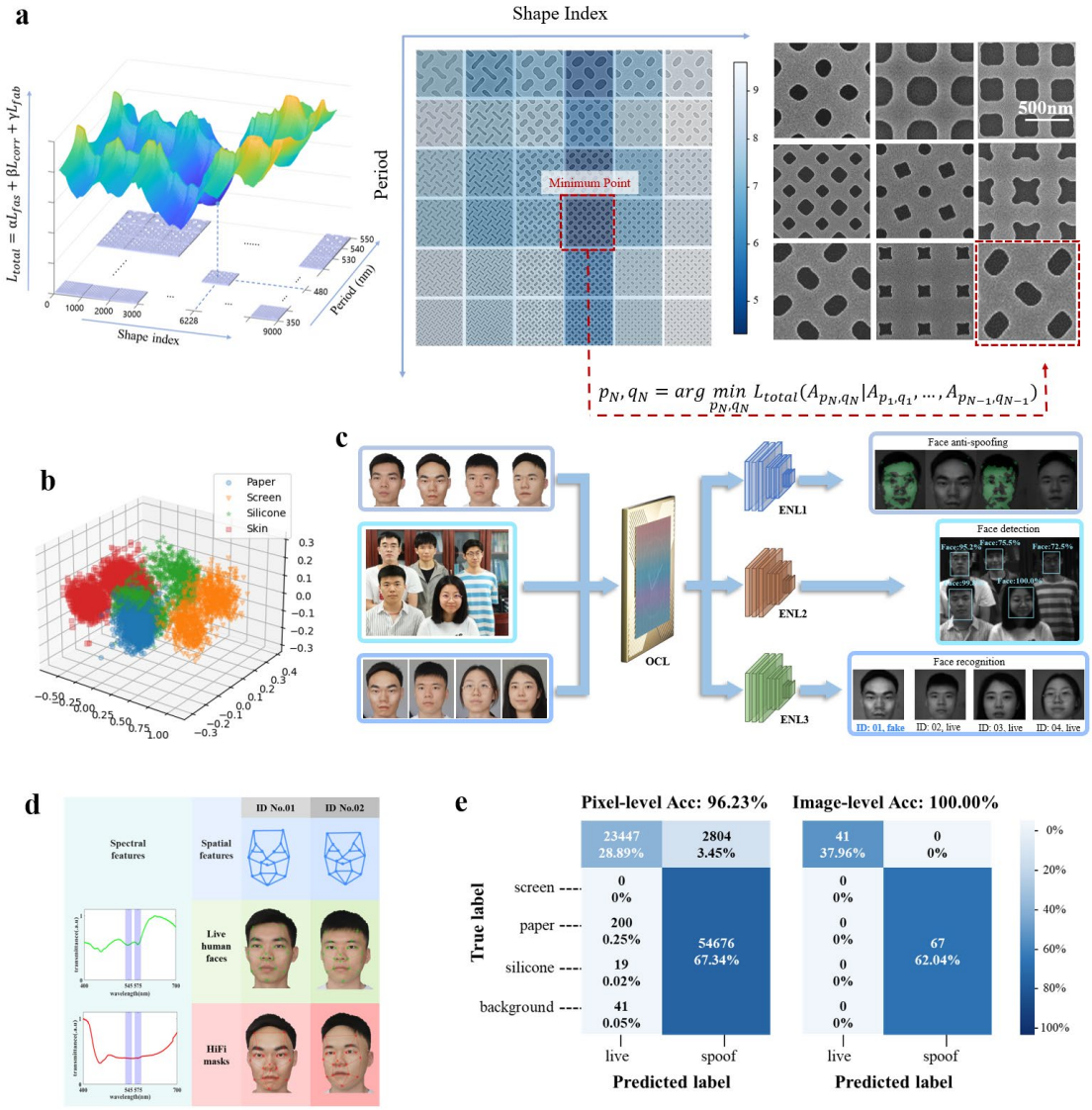


Fig. 1 Metasurface based SCNN chip can be used for multiple vision tasks related to face recognition. a, The GMTO algorithm is achieved by finding the minimum point of the designed loss function. **b,** Spectral feature extraction results of the OCL visualized by PCA. Live skin and three spoof materials are separated. **c,** The optical convolutional layer (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.

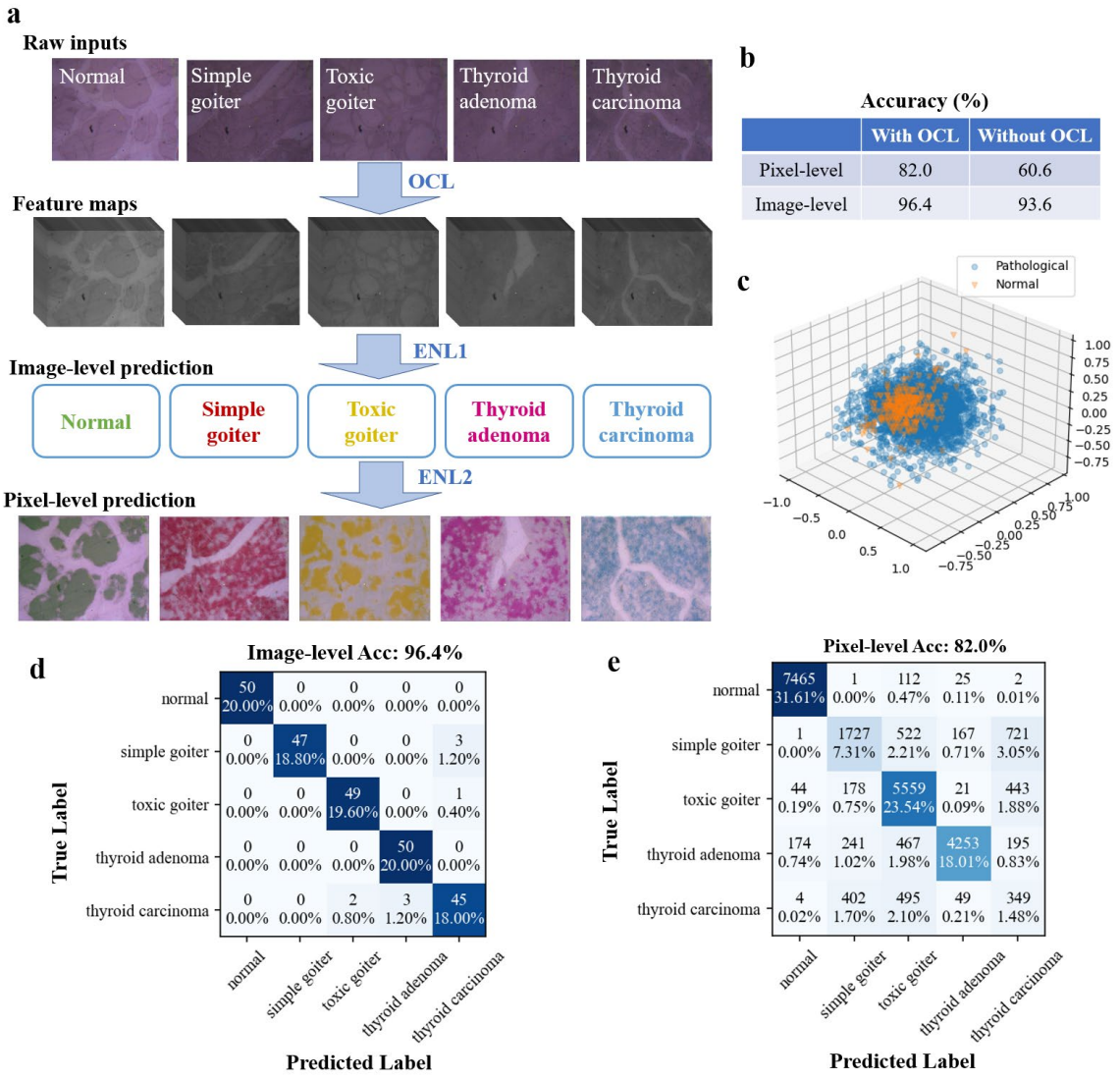


Fig. 2. Experimental results of thyroid histological section diagnosis by the Metasurface based SCNN. a, We exploit our SCNN to sense the raw datacube of thyroid histological section through a microscope. After the data are processed by the optical convolutional layer (OCL) and electrical network layers (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 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.

2. Add quantitative analyses of the proposed OCL. We describe the computing capabilities of the OCL in detail, and fully compare it with existing works in the revised manuscript of Table 1:

Table 2 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 digits 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**	**21.0 TOPS**	**5.3 TOPS/mm2**	√	√	√	**complex tasks in the real world: face anti-spoofing and disease diagnosis**

MMI: Matter Meta-Imaging

We expect that the revised manuscript can provide more in-depth explanations and bring an additional perspective for the advantages of the proposed Spectral Convolutional Neural Network.

To reviewer #3:

Comments: *Summary: This paper proposes a free-space optical dot-product engine for analog image pre-processing, combined with later digital CNN; it can realize high accuracy in 2 selected classification tasks. Experimental demonstration has shown its usage in image classification tasks.*

Response: Thanks for the time and patience in reviewing our manuscript. The comments are valuable for us to improve our work. We have realized that our work lacks some essential quantitative analysis and comparisons with other works. This may cause some misunderstanding or a misjudgment about our work. We have made major revisions to our manuscript and supplementary material, and the responses to the comments are listed below:

Comment 1: *The novelty and intellectual contribution are limited from the circuit/architecture/algorithm design sides. The engineering efforts in building such a demo are appreciated. Extensive designs and experimental demonstrations have been conducted on free-space optical dot-product engines, convolution engines, diffractive neural networks, etc. Some of them are even multi-layer, nonlinear, reconfigurable, and working with visible light, and can handle phase/polarization. A thorough comparison to prior free-space optics and integrated photonic accelerator designs is needed, especially on cost, efficiency, speed, throughput, size, reconfigurability, reliability, robustness, expressivity, etc.*

Response: Thank you for your insightful concern and guidance, which remind us that the device architecture and working principle are not descripted clearly enough. We have addressed the contributions and advances of the proposed spectral convolutional neural network (SCNN) more clearly in our revised manuscript and supplementary material. Besides, we added the comparison with prior integrated photonic accelerator designs in Table 1, which is also presented below.

Actually, to the best of our knowledge, we are the first to achieve on-chip optical computing utilizing natural light with in-sensor edge-computing capability. Our work is integrated optics as the device is totally on-chip rather than free-space optics. On-chip sensing and computing of natural broadband spectral images are the most important features of the proposed OCL. Indeed, some of the existing works have achieved multi-layer and nonlinear optical neural networks using free-space optics. However, these works based on free-space optics are incapable of edge-computing for portable terminals. Compared with free-space optics, our chip is very compact but it can still achieve very high computing density (about 5.3 TOPS/mm^2) and can be applied to edge devices. On the other hand, compared with on-chip integrated optics, our work does not rely on a coherent light source or fiber coupling. Moreover, our chip is integrated on the image sensor, thus empowering in-sensor computing capabilities. As most of the existing works can only perform simple tasks such as edge detection and handwritten digits recognition, while the proposed SCNN realize complex real-world tasks far beyond handwritten digits recognition. Last but not least, we have achieved mass production of the SCNN on a 12-inch wafer. We believe that the proposed SCNN could open a new practical in-sensor computing platform for complex vision tasks with matter meta-imaging (MMI) functions in the real world.

Table 3 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 digits 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**	**21.0 TOPS**	**5.3 TOPS/mm2**	√	√	√	**complex tasks in the real world: face anti-spoofing and disease diagnosis**

MMI: Matter Meta-Imaging

Comment 2: *The demonstrated optical convolution unit is equivalent to a Conv2d(in_channels=1, out_channels=9, kernel_size=1, stride=1, bias=False) or Conv2d(in_channels=1, out_channels=1, kernel_size=3, stride=3, bias=False) with fixed, quantized, noisy, positive-value weights. The expressivity of such a CONV layer is a concern, given that advanced real-world CV tasks require much more complicated DNN models. By replacing the first CONV of a DNN with this optical CONV, what are the overall impacts and system-level benefits?*

Response: Thank you for your detail comments, and we apologize for any confusion caused by our description. What we intended to claim is that, the input channel number of the demonstrated OCL is C rather than 1 (where C represents the spectral channels). We made revisions to the manuscript (page 6 lines 145-148) to describe the function of OCL more clearly. For the proposed spectral convolutional neural network, the OCL accepts 3D datacube (C-channel images) of two spatial dimensions and one spectral dimension as inputs rather than single-channel images. Accordingly, the OCL performs computing in the spectral domain, and the spectral sensing capability is one of the most important features of the proposed OCL. In our implementation, C is related to the

spectral bands. The quantitative analysis of C and computing speed have been added in Supplementary Note 2.

On the other hand, the weights of our OCL are not quantized and noisy. The OCL performs analog computing rather than digital computing. Moreover, for edge computing applications, our OCL is not designed to perform in-situ training. The fixed and positive-valued weights have show accuracies as high as 96.4% and 100% for pathological diagnosis and real-time face anti-spoofing at video-rate on such complex real-world tasks. These results indicate the fixed and positive values performs well even for complex real-word edge-computing applications. Besides, our ablation study demonstrated in the manuscript has further shown that the performance of the whole network will drop dramatically without the OCL. Taking the results of metasurface-based SCNN as example, "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%" (page 12 lines 268-270).

Finally, we also give a quantitative analysis (Supplementary Note 2) which indicates that our OCL can significantly reduce the computational burden of processing hyperspectral images, thus realize the in-sensor edge computing abilities which is impossible for previous hyperspectral imaging task. If we replace the OCL with a digital convolutional layer, then the digital layer has to process the multi-channel and high-resolution hyperspectral images, which is computationally expensive and brings great difficulties to edge computing. The OCL can reduce about 569.3M operations for processing a single hyperspectral image. It is in-sensor computing that provides a computing speed as high as 21.0 TOPS and a substantial reduction of 96% in data throughput, so that the computational load of the electrical backend can be significantly reduced. (see detail in Supplementary Note 2 and Table 1. Moreover, the exposure time of the CIS is relatively low (sampling rate is about 37 kHz) compared with high-speed photodetector (sampling rate can exceed 100 GHz). If we replace the CIS with PD array, the computing speed can be further improved to over 10^7 TOPS. In addition, to acquire such hyperspectral images, an extra hyperspectral camera is needed and high-performance electronic computing platforms, such as graphic process units (GPU) are inevitable for conventional strategies.

Comment 3: *The reconfigurability is big concern, how to enable reconfigurable weights? And what are the underlying trade-offs?*

Response: Thanks for your valuable questions and we have addressed this concern in our revised manuscript: "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. It is a tailored chip for a specific task for edge computing applications." (page 15 lines 353-355) The reconfigurable weights will significantly increase the system complexity and result in much lower integration. Especially, electrical computing provides much stronger training capabilities than existing optical neural networks. Therefore, the best way is to train and design the network by electrical computing, and we can utilize optical computing to greatly reduce the computing burden at inference. Moreover, the optical computing is performed in-sensor. The highly parallel vector inner-product is driven by the energy of input natural light and completed during the light field sensing process. In-sensor computing can greatly reduce the computing and storage burden of the downstream algorithms.

Comment 4: *The claimed reconfigurable kernel size sacrifices the efficiency by electronically summing the dot-product results, which might not be very efficient.*

Response: Thanks for the suggestions and a quantitative analysis is added to Supplementary Note7. We found that the computing burden of this electrical summing has little effect on the overall efficiency, and the detail analysis is as follows. Considering a convolutional kernel of shape $n \times n \times C$, at each spatial location, the kernel will perform n^2 dot-product operations and n^2 summing operations. Each dot-product operation requires C multiplications and $C-1$ summing operations. In all, we need $n^2(2C-1)+n^2 = 2n^2C$ operations. The final summing of the dot-product results only accounts for $\frac{n^2}{2n^2C} = \frac{1}{2C}$ of the computing burden. As C is usually a large number (For example, 216 as illustrated in Supplementary Note 2), the computing burden of the summing of the dot-product results can be neglected. Moreover, this summing operation can also be completed by binning during the readout process of the image sensor. Therefore, summing of the dot-product results shows little impact on the overall efficiency.

Comment 5: What is the robustness of such a system? Quantization, resolution, fabrication variation, thermal sensitivity, alignment sensitivity, signal-to-noise ratio, etc.

Response: Thanks for the synthetical comments. As our device is an integrated architecture, it has relatively high robustness. Based on the advantages of this integrated framework, we have achieved mass production by lithography in a standard semiconductor foundry on a 12-inch wafer. Therefore, it can be roughly predicted that the fabrication variation, thermal sensitivity, alignment sensitivity, and signal-to-noise ratio are very similar to a common commercial monochrome or RGB camera. Lastly, the OCL performs analog computing, the spatial resolution realized in this work is as high as 400×533 , since the SCNN provides the strategy of utilizing every single pixel to perform optical computing via CIS to achieve high computing density (Supplementary Note 2).

Comment 6: *The output feature map is 400×533×9, which is a large feature map. What is the data movement cost, bandwidth requirement, and system throughput?*

Response: Thanks for the valuable question and an analysis is added to Supplementary Note 2. Firstly, the OCL can reduce about 569.3M operations for processing a single hyperspectral image. It is in-sensor computing that provides a computing speed as high as 21.0 TOPS and a substantial reduction of 96% in data throughput, so that the computational load of the electrical backend can be significantly reduced. Secondly, the OCL can reduce hyperspectral images into 9-channel feature maps, which is a 96% reduction in data movement cost, bandwidth requirement, and system throughput for transferring hyperspectral images. These computing and data compression capabilities are exactly the advantages of the proposed OCL. Finally, as illustrated in the manuscript, our whole chip is integrated on a monochrome image sensor. The data movement cost, bandwidth requirement, and system throughput are all similar to a common monochrome or RGB camera, which can be applied to edge computing.

Comment 7: *How does that compare to standard CNN taking the RGB image as inputs?*

Response: We have added related descriptions in our revised manuscript (page 12 line 273). RGB sensor cannot provide spectral sensing capabilities. In the RGB color space, the live and spoof faces (or sections of tissues with different diseases) are not distinguishable. Therefore, common RGB sensor cannot achieve such tasks.

Comment 8: *If the main advantage is from the MPCF in processing multiple spectrums of the image, how does it compare to metasurface-based DONN that can sense and process other dimensions, e.g., polarization and phase?*

Response: We really appreciate the reviewer's suggestion of using metasurface-based modulators. We have adopted this suggestion and conducted several experiments on the metasurface-based spectral filters according to your comments. The results are presented in the revised manuscript, which helped us to further enrich the experiments and greatly improve the manuscript. Two new figures (Fig. 2 and Fig. 3) are added to demonstrate the metasurface-based SCNN results, which is also shown below. The comparison between metasurface-based and pigment-based SCNN is describe on page 15 lines 333-343 of the revised manuscript. 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, resulting in higher spectral resolution 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 (Refs 43-45). 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. In conclusion, it can be predicted that SCNN chips have great potential in various specific terminal applications.

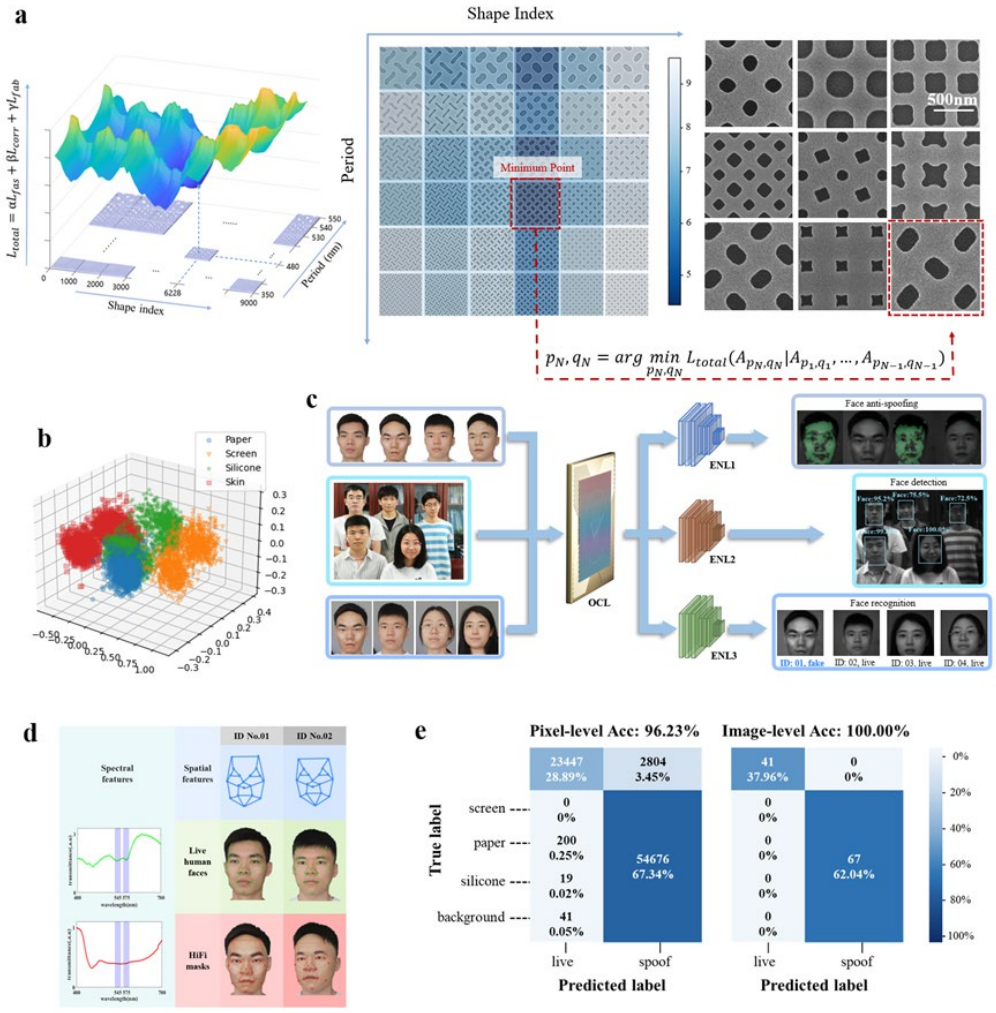


Fig. 3 Metasurface based SCNN chip can be used for multiple vision tasks related to face recognition. a, The GMTO algorithm is achieved by finding the minimum point of the designed loss function. **b,** Spectral feature extraction results of the OCL visualized by PCA. Live skin and three spoof materials are separated. **c,** The optical convolutional layer (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.

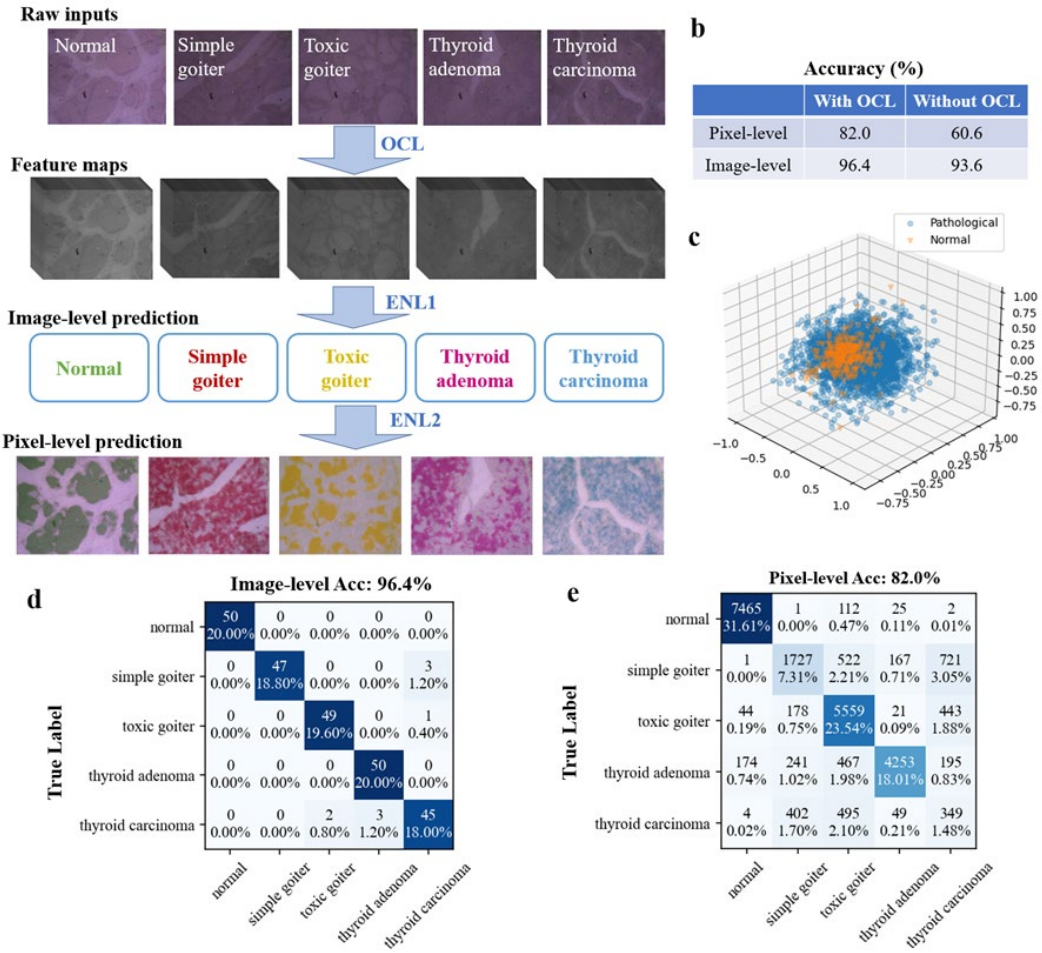


Fig. 4. Experimental results of thyroid histological section diagnosis by the Metasurface based SCNN. a, We exploit our SCNN to sense the raw datacube of thyroid histological section through a microscope. After the data are processed by the optical convolutional layer (OCL) and electrical network layers (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 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.

Comment 9: *By checking the code provided, the images are preprocessed and stored as 9-channel input features, and pass through a very deep CNN and ResNet, which makes the initial optical CONV almost meaningless. Why not just input the raw features from the sensor to the used large digital full-precision CNN running on GPU? The claim that it is very efficient on edge devices without the need of GPU is not very justified.*

Response: Thanks for your important feedback, which remind us that the advantages and significance of the proposed device are not demonstrated enough. We have added related content in the revised manuscript (page 16 lines 358-365). The 9-channel feature maps are exactly the raw

features from the sensor, i.e., the outputs of the OCL. Here, the OCL performs in-sensor computing, which capturing and transferring the natural hyperspectral images into feature maps of 400×533×9. This is exactly the advantage of the proposed OCL. If we remove the OCL, we can only get grayscale images from a common monochrome camera which does not contain any spectral features. Furthermore, our supplementary video has demonstrated the capabilities of our SCNN without GPU. Specifically, the SCNN performs in-sensor computing, which can reduce about 569.3M operations for processing a single hyperspectral image. The OCL is in-sensor computing that provides a computing speed as high as 21.0 TOPS and a substantial reduction of 96% in data throughput, so that the computational load of the electrical backend can be significantly reduced. (Detailed analysis has been added in Supplementary Note 2). Thus, the proposed SCNN opens a new practical in-sensor computing platform for complex vision tasks in the real world.

Indeed, a very deep CNN can also process hyperspectral images on GPU, but we cannot integrate it on edge devices. As is described in the revised manuscript: “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 computing speed as high as 21.0 TOPS and a substantial reduction of 96% in data throughput (Supplementary Note 2). Therefore, 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.” (page 16 lines 358-368)

To reviewer 4:

Comments: *The authors in this work proposed an integrated spectral convolutional neural network (SCNN) framework with in-sensor computing capability to detect visual information in broadband natural incoherent light. Thus, the computing speed can be improved and the energy efficiency is enhanced. The results are interesting. The authors are suggested to address my concerns before the manuscript being published.*

We appreciate the referee's thorough review, accurate summary, and supportive feedback on our work. The comments are essential for us to improve the manuscript. We have realized the lack of quantitative analysis and comparisons about the proposed optical convolutional layer (OCL). We have added these comparisons in the revised manuscript and Supplementary Material. Below, we address each of the raised concerns in detail.

Comment 1: *What is the computing speed and power consumption of the proposed SCNN chip, is it ahead of existing architectures?*

Response: Thanks for your question, which helped us to realize the lack of quantitative analysis in our manuscript. We have added the analysis and discussion in the revised manuscript (page 16 line 364) and Supplementary Note 2. The computing speed of the proposed OCL is about 21.0 TOPS and the computing density is about 5.3 TOPS/mm^2. It can exceed most of the existing architectures while providing sensing capabilities. The detailed comparison can be found in Table 1, which is also presented below. "For the OCL, the computing speed and power consumption depends only on the exposure time and the power of the CIS, empowering ultrafast optical computing at high energy efficiency." "Although CIS is relatively slow (sampling rate is about 37 kHz) compared with the commonly used high-speed photodetector (sampling rate can exceed 100 GHz), we still achieve considerable computing speed and density compared with existing photodetector-based works because CIS has high integration and can take full advantages of space division multiplexing. If we replace the CIS with PD array, the computing speed can be further improved to over 10^7 TOPS. 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, which still has the computing speed of over 10^{13} OPS (see Supplementary Note 2 for details). Thus, the proposed SCNN opens a new practical in-sensor computing platform for complex vision tasks with MMI functions in the real world."

Table 4 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 digits 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**	**21.0 TOPS**	**5.3 TOPS/mm2**	√	√	√	**complex tasks in the real world: face anti-spoofing and disease diagnosis**

MMI: Matter Meta-Imaging

Comment 2: *The authors noted that CNNs require significant computational resources. Is SCNN more lightweight? Please provide a quantitative analysis.*

Response: Thanks for the important suggestion, which remind us that the quantitative analysis of the proposed device are not demonstrated enough. Detailed analysis has been added in Supplementary Note 2. Our SCNN performs in-sensor computing, which can reduce about 569.3M operations for processing a single hyperspectral image. It is in-sensor computing that provides a computing speed as high as 21.0 TOPS, so that the computational load of the electrical backend can be significantly reduced Thus, the proposed SCNN opens a new practical in-sensor computing platform for complex vision tasks in the real world.

Comment 3: *In Figure 3, how is Acc calculated? Are there other more diverse evaluation metrics available? If so, please provide an analysis.*

Response: Thanks for the valuable feedback. Fig. 3 provides the confusion matrix of the SCNN on disease diagnosis applications. The disease diagnosis is a 5-class classification task. Assuming that the confusion matrix $M = \{m_{ij}\} \in R^{5\times5}$, each row of M represents a true label and each column of M represents a predicted label. m_{ij} represents the number of testing samples that have true label i and are predicted to be j. Therefore, the diagonal elements indicate the number of samples that were correctly categorized. The Acc (accuracy) metric is calculated as $\frac{tr(M)}{sum(M)}$, which indicates

the overall classification accuracy. For the classification task, the other commonly used evaluation metrics besides accuracy are precision (calculated as $\frac{m_{kk}}{\sum_{i=1}^{5} m_{ik}}$) and recall (calculated as $\frac{m_{kk}}{\sum_{j=1}^{5} m_{kj}}$) for each class k. Here we provide the precision and recall for each class in the table below, and we have added these additional analysis to Supplementary Note 8.

	Normal	**Simple goiter**	**Toxic goiter**	**Thyroid adenoma**	**Thyroid carcinoma**
Precision	100%	88.89%	98.63%	98.77%	96.25%
Recall	100%	96.00%	90.00%	100%	96.25%

Comment 4: *In the tasks of Histological Section Diagnosis and Face Anti-spoofing, how do existing methods perform? What are the advantages of the method proposed by the authors compared to existing methods? Please provide a detailed explanation.*

Response: Thanks for the constructive suggestions. As is described on page 15 lines 346-351, the main advantages of our proposed OCL are capturing and feature extracting of natural hyperspectral images. If we do not use the OCL, i.e., capturing images using a common monochrome camera and following the regular neural network training process, the results are shown in Fig. 3 and Fig.4 as "Without OCL". The performance is much worse than using the OCL because OCL provides spectral sensing capabilities. Taking the results of metasurface-based SCNN as example, "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%" (page 12 lines 268-270).

On the other hand, if we complete the whole process by capturing data using a hyperspectral camera and implementing all neural network layers on the electrical computing platform of GPU, 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. It is not practical in real-time applications. And as demonstrated above, the storing and processing cost of a hyperspectral image on an electrical computing platform require high-performance GPU, which cannot be integrated on edge devices. This is because that GPU as the high-performance electronic computing platforms with large size, high energy consumed, and high cost cannot meet the requirements of edge devices with limited computing capabilities.

Different from the conventional strategies, our OCL is in-sensor computing providing a computing speed as high as about 21.0 TOPS, which makes it possible to process hyperspectral images using only a few extra digital neural network layers of edge devices as experimentally demonstrated in the manuscript, which empowers edge devices with both sensing and computing capabilities for various real-world complex vision tasks.

Comment 5: *In this manuscript, a neural network chip with photoelectric hybrid architecture is proposed, a comparison table including key parameters with existing on-chip neural network should be provided.*

Response: Thank you very much for the suggestion, which is valuable for us to improve the manuscript. The comparison with existing on-chip methods is listed above (Table 1). The table is also added to the manuscript.

Comment 5: *If available, please provide both quantitative and qualitative comparisons with existing methods.*

Response: Thanks for the valuable suggestion. As is mentioned above, abundant revisions have been made to the manuscript. The comparison with existing on-chip methods is listed as Table 1 in the revised manuscript and related analysis is also added in Supplementary Note 2.

To reviewer #1:

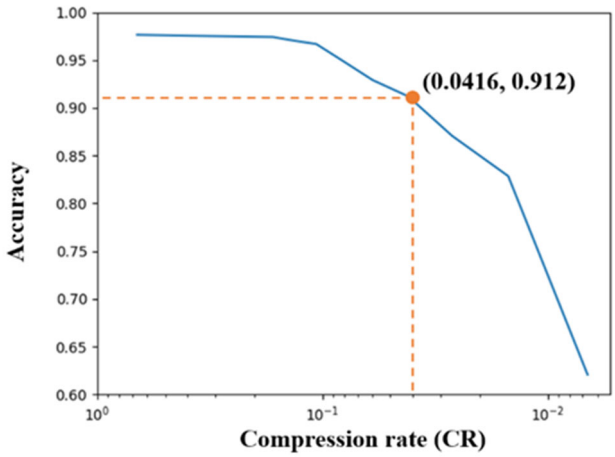
Comment: *Thank you to the authors for responding to the review comments and concerns. Thank you for the effort made to address all the comments; however, the authors' rebuttal does not adequately resolve two remaining points. The argument in the revised manuscript has the following problems, especially 1 question the validity of the claims*

Response: We sincerely thank you for your patience and time in reviewing our revised manuscript and giving further valuable questions. Your comments are helpful for us to further improve the manuscript. Below, we also address each of the raised concerns in detail.

Comment 1: *The calculation of the number of operations performed is suspect. First, the authors should first demonstrate that the described compression rate of 4.16% results in no loss performance for the implemented tasks. Second, the throughput of 2NC/T does not use an appropriate sampling time. The camera sensor as detailed is the CS235MU with a maximum full frame-rate of 165.5fps. From this readout, it is inappropriate to use T=0.027ms, and instead T = 6ms is more appropriate. The comparison to a high-speed PD array for a throughput of 10^7 TOPS is also inappropriate as the corresponding data rate would be 40000TB/S, and that doesn't even include a host of other problems with light intensity and electronics constraints.*

Response: We sincerely appreciate the question. We realized that some descriptions in the article lack further explanations and support. Therefore, we have added additional explanations, references, and experiments. We have also revised the description of the manuscript to demonstrate our claims more rigorously.

For the compression rate of 4.16%, we have added experimental results to explain that such a compression rate is attainable for the implemented tasks, such as face anti-spoofing. These results are added to Supplementary Note 2: "In previous works, hyperspectral sensing with 601 sampling points (from 450~750nm at 0.5nm intervals) is realized using only 25 spectral filters[1,5]. That is, the compression rate (CR) is about 4.16%. Other works, such as Ref. 4 (CR=5%) and Ref. 11 (CR=1%~10%) have also illustrated that similar compression rates can effectively reserve the spatial and spectral features. To further study the impact of CR in experiments, we test the liveness detection performance under different CRs using a snapshot hyperspectral camera with spectral filters of as many as 400, which is developed in our previous works (Ref. 38). The camera is deployed to capture spectral pixels of live human skin and spoof masks. We adjust the CR by changing the number of spectral filters valid in one spectral pixel. Then, we utilize the support vector machine (SVM) algorithm to perform classification. The classification accuracy is reported in Supplementary Fig. 1. Under the CR of 4.16%, we can still realize a single-pixel classification accuracy of 91.2% by SVM. The accuracy of 91.2% is close to the accuracy of 96.2% shown in Fig. 2e of our manuscript, we consider that this little difference lies in that our previous camera is not specially optimized for liveness detection and the performance of a single-layer SVM is not as good as that of a multi-layer DNN demonstrated in Fig. 2c."



Supplementary Fig. 1. The liveness detection accuracy under different CRs.

For the sampling time, we have also added a more rigorous description in Supplementary Note 2. For the metasurface-based sensor, "The CIS used in our experiment was a Thorlabs CS235MU equipped with a Sony IMX249 sensor. The minimum exposure time is 0.034 ms. When the sensor completes the exposure, the computation of OCL is also complete. Therefore, $N = 480 \times 366, C = 216, T = 0.034ms$, then the theoretical maximum computing speed of the OCL is about 2.2 TOPS. However, from the perspective of the overall system we have implemented, the maximum full-pixel (480×366) frame rate achieved on our laptop computer (Thinkpad X1) is 116.8 frames per second (FPS). In this way, $T = 8.65ms$ and the average computing speed of OCL is calculated to be 8.7 GOPS. It is worth noting that the computing speed of OCL only depends on the imaging speed of the CIS because the OCL performs in-sensor computing. The OCL is designed for real-world vision tasks and the bottleneck stays in the vision sensor itself. No matter how fast the imaging speed is, the OCL can ensure that the computing is completed once an image is captured. 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 tasks. Moreover, in our implemented system, the frame rate of 116.8 FPS is already sufficient to complete most real-world computer vision tasks and provide spectral sensing abilities for edge devices. By further increasing the system's transmission bandwidth (for example, use PCIe instead of USB for data transfer), the actual average computing speed of OCL can be pushed to the theoretical maximum computing speed." For the pigment-based sensor, "The minimum exposure time is $0.027ms$. Therefore, $N = 1200 \times 1098, C = 216, T = 0.027ms$, then the theoretical maximum computing speed of the OCL is about 21.0 TOPS and the OCL can reduce the storage requirement to 7.3MB. Similarly, limited by the sensor readout time and USB 3.0 transmission speed, the maximum full-pixel (1200×1098) frame rate achieved on our laptop computer (Thinkpad X1) is 30.2 FPS and the average computing speed of OCL is calculated to be 17.2 GOPS." We have also added some qualifications to the description of the computation speed in the manuscript on lines 332-336, making it more accurate: "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. The theoretical maximum computing speed of OCL is about 21.0 tera operations per second (more detailed analysis can be found in Supplementary Note 2)."

For the claims about the PD, the state-of-the-art commercial products are capable of reaching

100GHz (such as the BPDV412xRv released by Coherent Corp.). Besides, existing works have adopted relatively high-speed PD to increase computing speed (such as the 12 GHz PD adopted in Ref. 26 and the 50 GHz PD adopted in Ref. 27). The sampling rate of PD can be 10^5~10^6 times faster than CIS. As the primary factor limiting our computing speed is the sampling rate of the detector, if a high-speed detector such as PD is used to increase the sampling rate to 20 GHz, the theoretical maximum computing speed of our OCL is predicted as 10^7 TOPS. However, as the reviewer pointed out, great challenges stand in the system data throughput and PD array integration, this claim of the 10^7 TOPS computing speed is disputable. Therefore, we have deleted the corresponding quantitative description in the manuscript based on the reviewer and revised the claim on lines 346-348: "If we replace the CIS with high-speed PD array, there is still great potential for improvement in computing speed."

Comment 2: *In the text, the kernel sizes are written: "its kernel size n and number of kernels $K = k2$ can be reconfigured as well as $k \cdot n$ is fixed to the size of the OCU," along with "Therefore, OCL has KK convolutional kernels of size $n \times n$ and stride $n \times n$." While this is an accurate description of the system, it obfuscates the point that the sensor size is fixed and the input light is filtered by the SCNN without the possibility for additional spatial mixing, as would be expected for a convolutional layer. As the primary successful implementation of the SCNN in the manuscript was using nine designed spectral filters, I believe an emphasis of the SCNN as primarily a high-speed customizable hyper-spectral imaging method would only serve to strengthen the manuscript.*

Response: Thanks for the suggestion, your advice is very valuable for us. We took the suggestion and added several revisions to our manuscript. After fabrication, the size of the OCU is fixed. When the kernel size is not 1, the stride cannot be 1 either. Therefore, the proposed OCL is actually a strided convolutional layer. To address this point and avoid any confusion that might be caused by our former description, we have revised the manuscript and added additional explanations on lines 139-146: "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. 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". Just as you pointed out that the SCNN is indeed a high-speed customizable hyperspectral imaging method. We have also emphasized this perspective on lines 79-82: "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."

To reviewer #2:

Comment: *I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.*

Response: We sincerely appreciate your time and patience in reviewing our revised manuscript. As listed in this response letter, we have provided a point-by-point response to every newly raised concern. Several revisions have been made to further support the claims in the manuscript.

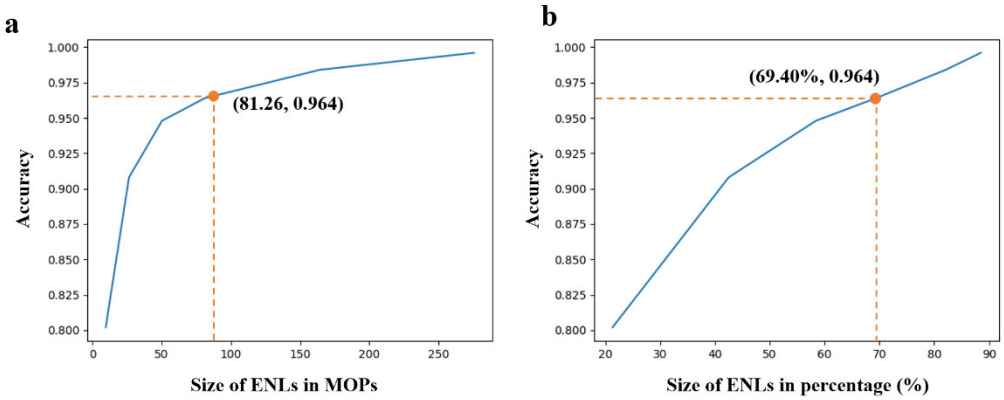
To reviewer #3:

Comment 1: *For each hyperspectral image, the number of operations saved is 569.3M operations, which is only the operations for a standard 3x3 Conv2d in a DNN. Compared to the rest of the network layers, it is not a significant computation reduction.*

Response: Thanks for the valuable question. The optical convolutional layer (OCL) works as the first layer of the whole optoelectronic neural network. For the pigment-based OCL, indeed it can be regarded as a Conv2d in a DNN. However, the inputs are hyperspectral images rather than common RGB images. As is mentioned in Supplementary Note 2, the number of input channels can be regarded as 216. Therefore, the input shape *(channels, height, width)* is *(216, 400, 533)* and the required operations for this Conv2d is 569.3M. Note that the same Conv2d for RGB images with input shape *(3, 400, 533)* only requires 7.9M operations. Therefore, the operations provided by our OCL are equivalent to about 72 RGB-based Conv2d layers.

In particular, powered by the feature extraction and data compression of this OCL, the computational load and data throughput of the following electrical network layers (ENLs) can be greatly reduced, which is the main advantage of our framework for edge computing. Namely, we can reach a considerable performance using small-size ENLs in common cases with the help of the proposed OCL. For example, the disease diagnosis results demonstrated in Fig. 3e are achieved by small-size ENLs. The SCNN is an optoelectronic neural network, although the optical part only forms the first convolutional layer, it accounts for 30.60% of computing operations while the other 69.40% of operations are completed by the reduced ENLs.

We fully understand the reviewer's concern that the mentioned 569.3M operations may account for only a small part of the entire network and we thank you again for the valuable question. To fully address your concern and demonstrate the role of OCL, we have added several additional quantitative analyses to Supplementary Note 9, which is also displayed below:

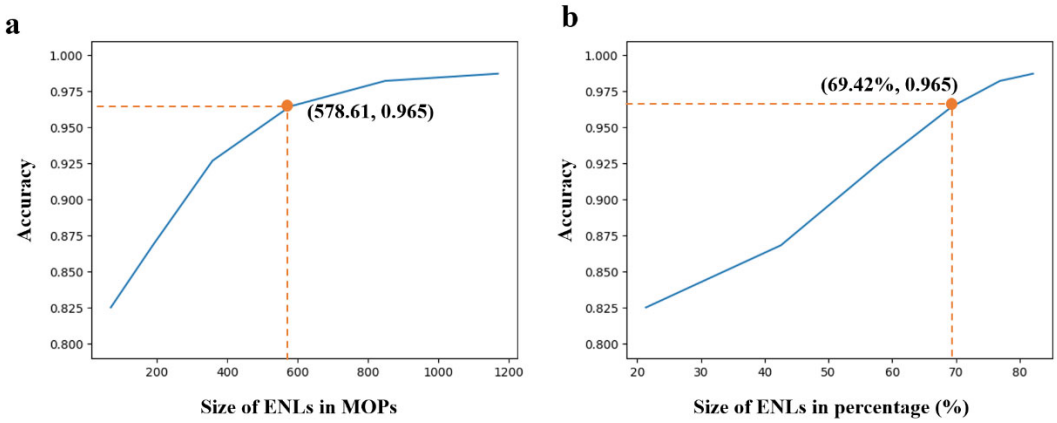


Supplementary Fig. 12. The image-level classification accuracy v.s. the size of ENLs on metasurface-based SCNN. a, The size of ENLs is represented by MOPs. **b,** The size of ENLs is represented by the percentage of computational load in the whole SCNN.

The final classification performance is related to both the OCL and ENLs. To study the influence of

ENLs on classification performance, we change the size of ENLs and test the final accuracy of the metasurface-based SCNN on the disease diagnosis task. The size of ENLs is represented by both MOPs (Supplementary Fig. 12a) and by its percentage of computational load in the whole SCNN (Supplementary Fig. 12b). If we adopt the network described in Supplementary Fig. 7, the ENLs need 81.26 MOPs, which account for 69.40% of the whole SCNN and the other 30.60% operations are completed by the OCL, and the final accuracy is 96.4%. Supplementary Fig. 12a also shows that the classification accuracy drops sharply when the size of ENLs is less than 50 MOPs and grows slowly when the size of ENLs is greater than 50 MOPs. If we adjust the size of ENLs to 50.35 MOPs, which only account for 58.42% of the whole SCNN, we can still achieve an accuracy of 94.8%. The ENLs can have large sizes to achieve a high-performance super-resolution task, such as the ENLs adopted in the pixel-level liveness detection tasks, but it is unnecessary in most computer vision tasks.

These results indicate that, although larger ENLs can lead to better results, we can still reach a considerable performance using small-size ENLs because the OCL can provide powerful in-sensor feature extracting capabilities. Similar results can also be achieved by the pigment-based SCNN, as shown in Supplementary Fig. 13.



Supplementary Fig. 13. The image-level classification accuracy v.s. the size of ENLs on pigment-based SCNN. a, The size of ENLs is represented by MOPs. **b,** The size of ENLs is represented by the percentage of computational load in the whole SCNN.

Comment 2: *The claimed 10^7 TOPS using PD arrays is not valid, as no data movement solutions can support such a data readout rate.*

Response: Thanks for the suggestion. The state-of-the-art commercial products are capable of reaching 100GHz (such as the BPDV412xRv released by Coherent Corp.). Besides, existing works have adopted relatively high-speed PD to increase computing speed (such as the 12 GHz PD adopted in Ref. 26 and the 50 GHz PD adopted in Ref. 27). The sampling rate of PD can be 10^5~10^6 times faster than CIS. As the primary factor limiting our computing speed is the sampling rate of the detector, if a high-speed detector such as PD is used to increase the sampling rate to 20 GHz, the theoretical maximum computing speed of our OCL is predicted as 10^7 TOPS. However, as the reviewer pointed out, great challenges stand in the system data throughput and PD array integration, this claim of the 10^7 TOPS computing speed is disputable. Therefore, we have deleted the

corresponding quantitative description in the manuscript based on the reviewer and revised the claim on lines 346-348: “If we replace the CIS with high-speed PD array, there is still great potential for improvement in computing speed.”

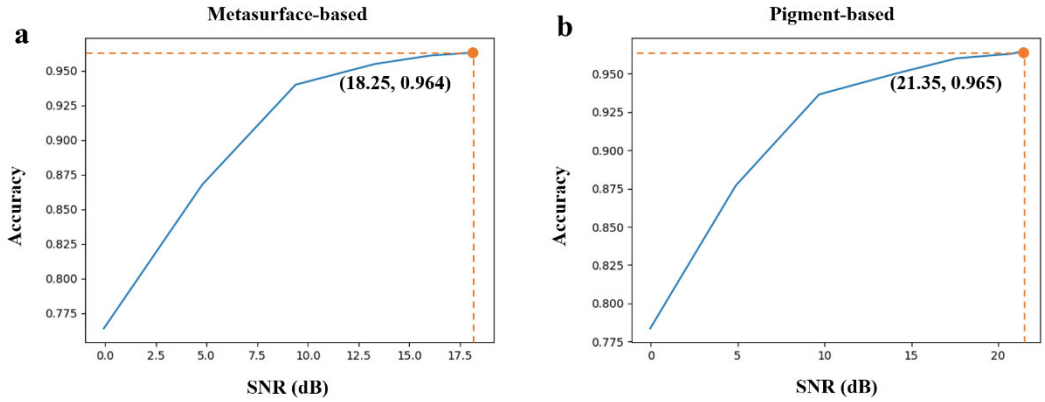
Comment 3: *The reconfigurability concerns of this fixed processor are not addressed. A technological solution to enable reconfiguration is required. It has nothing to do with training, just for other functionality to use this device. Otherwise, fixed functionality” should be put in the title.*

Response: We sincerely thank you for the advice. We also fully understand the reviewer’s concern that the optical part cannot be reconfigured. Therefore, we have revised the manuscript to further emphasize this point based on your advice on lines 317-321: “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 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”.
On the other hand, the proposed SCNN is an optoelectronic neural network. As the OCL cannot be reconfigured, the SCNN can be adjusted by changing the electrical part. As demonstrated in the manuscript on lines 26-28: “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.” Therefore, we do not put “fixed functionality” in the title. Instead, we further clarify the point of reconfiguration in the manuscript as mentioned above to avoid any misunderstandings.

Comment 4: *The authors claimed the weights are not quantized and noisy, which is not true. The fixed weights are from fabrication, it has to have certain precision and process variation.*

Response: We sincerely apologize for misunderstanding your previous question. We thought the noise you mentioned referred to the errors caused by weight quantization in digital computing. Your suggestion is very valuable. We have measured the noise level and further studied the influence of the noise according to your suggestion. The results are reported in Supplementary Note 10, which are also presented below:

“In the proposed SCNN framework, all of the OCUs are regarded to be identical. However, due to the fabrication precision, readout noise, quantization error from analog-to-digital conversion, etc., there are certain variations between these OCUs, resulting in the noisy output of the OCL. The SNRs of metasurface-based and pigment-based OCL outputs are measured to be 18.25 dB and 21.35 dB. The pigment-based OCL is taped out on a 12-inch wafer by a standard semiconductor lithography process and can reach a good consistency. Therefore, the SNR is relatively high. The metasurface-based OCL is fabricated by electron beam lithography (EBL), the fabrication precision can cause certain differences between different meta-atom units, thus having relatively low SNR.



Supplementary Fig. 14. The influence of SNR on the image-level disease diagnosis task. a, The accuracy-SNR curve of the metasurface-based SCNN. **b,** The accuracy-SNR curve of the pigment-based SCNN.

To further study the impact of OCL noise on the final performance, we add different levels of noise to the OCL outputs by simulation and test the final classification accuracy. The results are shown in Supplementary Fig. 14. The results indicate that the SCNN can maintain relatively high performance (over 93% accuracy) when SNR > 10 dB and the performance drops dramatically when SNR < 10 dB. As the SNRs of our OCL outputs are 18.25 dB and 21.35 dB, the SCNN can maintain a relatively high performance of over 96% accuracy, which indicates the impact of noise for the proposed structures on the final performance is relatively limited."

Comment: *The codes do not contain much photonic analog part. It is a pure digital CNN training code.*

Response: Thanks for the comment. We have added the codes for the photonic part to the GitHub repository for the revised manuscript and updated the Code Availability section: "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 https://github.com/rao1140427950/scnn_mpcf. Other algorithms and methods are included in this published article (and its supplementary information files)."

To reviewer #4:

Comment: *This manuscript has been completely revised based on the recommendations made. I recommend accepting it.*

Response: We sincerely thank you for your acknowledgment of our work. The comments and advice you provided earlier have been immensely helpful for us in improving the manuscript. Thank you again for your time and patience.

To reviewer #1:

Comment: *Thank you to the authors for the revisions in response to the previous round of reviews. One concern is that the paper makes claims that overstate what has actually been shown and the text needs to be toned down. This also seems to have been reflected in another review report. The authors have made changes to remedy some of these statements, but further changes are needed: Specifically, it is necessary for the authors to display the actual demonstrated performance of the experimental device (17.2 GOPS) and corresponding compute density in Table 1. The current number displays a potential performance (21.0 TOPS) that cannot be realized with the hardware used in the manuscript. This is misleading. Please update Table 1, as described above, to reflect what you actually show in the paper.*

Response: We sincerely thank you for the advice and suggestions. We understand your concern that the cost of data readout and transfer should be considered at the whole system level. However, the proposed OCL will complete the computing once the CIS has completed the exposure. The computing itself is completed before the data readout. Thus, the computing speed of OCL is only determined by the exposure time. As the minimum exposure time is 0.027ms for the fabricated pigment-based sensor, the computing speed of OCL can indeed attain 21.0 TOPS for the physical implementation.

For practical applications of real-world vision tasks, it usually does not require a particularly high frame rate at the whole system level. Thus, the OCL has an adaptive computing speed based on the imaging speed of the CIS. Accordingly, the average computing speed is reduced to 17.2 GOPS with a frame rate of 30.2 frames per second for practical vision tasks. It should be noted that the true advantage of our in-sensor OCL is that the computing speed can always satisfy the requirements of real-world applications. This is because the OCL can ensure that the computing is completed once an image is captured. The faster the image is captured, the faster the computing speed is.

Therefore, we have further revised the manuscript to avoid disputed claims about the computing speed. Specifically, we deleted the quantitative description of the computing speed in the main text (lines 108-109, 342-344) and Table 1. Instead, we address that the SCNN has adaptive computing speed and can always meet the imaging speed, and revised the description on lines 107-113 to reflect the true feature of our OCL “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.”

To reviewer #2:

Comment: *I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.*

Response: Thank you again for your time and patience in reviewing our manuscript.

To reviewer #3:

Comment: *I agree with Review 1 that the proposed SCNN is a customized (in the sense of fixed function after fab) spectral imaging/preprocessing method to collect information from multiple spectrums. It is not questionable and is better than collecting only RGB channels.*

Response: We truthfully thank you for the comments and we sincerely appreciate your acknowledgment that our proposed SCNN is better than collecting only RGB channels. The SCNN is indeed a customizable hyperspectral imaging method designed and implemented from the perspective of a neural network. we have addressed this statement on lines 84-87: "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."

Comment: *However, selling this chip (3x3/1x1 conv) as an edge NN accelerator that speeds up the whole NN system will have a lot of problems in speed, data movement, reconfigurability, etc.*

Response: Thank you again for the valuable advice. Actually, the computing speed and data throughput of the proposed OCL will have little impact on the performance of the whole system. Because the OCL performs in-sensor computing. We have further revised and explained the computing and data reduction ability of the proposed OCL on lines 107-113: "In this framework, the in-sensor 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." Besides, the demo videos recorded in real-world applications, which are provided as Supplementary Movies 1~2, have also shown that the proposed SCNN can achieve high performance in real-time on a notebook computer without any GPU.

Although the OCL is fixed after fabrication, the electrical backend can be reconfigured. We have demonstrated this feature on lines 74-77: "Hybrid optoelectronic computing hardware with an OCL and a reconfigurable electrical backend is employed to leverage optical superiority without sacrificing the flexibility of digital electronics". The capability of reconfigurable electrical backend is also shown by experiments on lines 27-30 that "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." Besides, we have addressed the explanation and solution of such a fixed OCL on lines 323-330: "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 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."