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

Kaiyu Cui[#*1], Shijie Rao[#1], 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], and Shengjin Wang[1]

[#] These authors contributed equally to this work

[*] Correspondent authors: kaiyucui@tsinghua.edu.cn, yidonghuang@tsinghua.edu.cn

Affiliation: [1]Department of Electronic Engineering, Tsinghua University, Beijing, China

[2]Beijing Seetrum Technology Co., Beijing, China

Supplementary Note 1. Comparison between SCNN and on-chip snapshot hyperspectral imaging.

On-ship snapshot spectral imaging (SSI) strategy needs to design a universal SSI chip for arbitrary spectrum reconstruction. It is based on compressive sensing theory. Therefore, SSI usually needs dozens to hundreds of different optical filters such as metasurface or Fabry-Pérot structures to get more compressive measurements and thus obtain higher spectrum reconstruction precision. Its applications focus on measuring spectra.

Take face anti-spoofing (FAS) as an example. SCNN is much more effective and practical. The comparison between our previous SSI-based method[1,2] and SCNN are listed below.

Supplementary Table 1: Comparison between SSI-based and SCNN-based FAS

	SSI[1,2]	**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	**×**	**√**

SCNN aims at designing an application-oriented chip for real-world computer vision tasks based on optical neural network (ONN). It provides an in-sensor computing and non-reconstruction spectral imaging method for the final target of the downstream task. SCNN can use minimal metasurface units by just extracting the spectral features for specific applications. This provides an ONN-based approach for hyperspectral sensing tasks, effectively avoiding the need for as many metasurface units as possible for high precision spectral reconstruction. Further, fewer kernels enable higher feature compression capability, higher spatial resolution, and extremely lower computing costs

for ENLs for the ONN with an optoelectronic framework. Generally, SSI systems need massive computing resources to complete the reconstruction procedure. Powered by modern artificial neural network technics and high-performance computing devices such as GPU, SSI systems are barely to achieve video-rate[3,4]. However, SCNN can reach video-rate easily even on a common laptop CPU, which empowers edge computing for terminal devices with limited computing capabilities.

Supplementary Note 2. Computing speed of optical convolutional layer.

The speed of analog computing cannot be simply quantified by operations per second (OPS). However, for a comparison with digital computing, we provide a method for calculating the equivalent OPS of our analog OCL based on the properties of our device. Notably, the computing speed is mainly limited by the pixels and the exposure time of the CMOS image sensor (CIS).

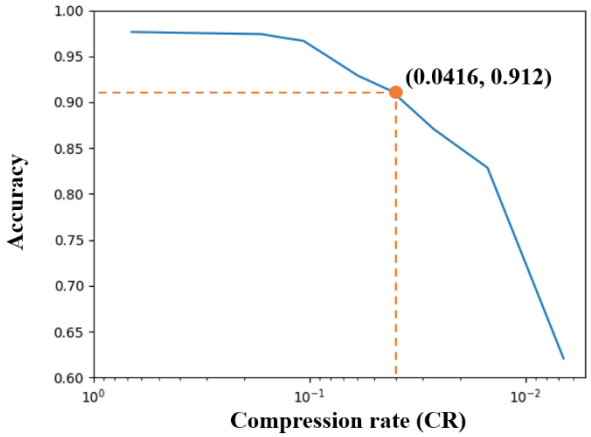
Considering a convolutional kernel of shape $n \times n \times C$, at each spatial location, the kernel will perform n^2 dot-product operations and n^2 summing operations. Each dot-product operation requires C multiplications and $C-1$ summing operations. In all, we need $n^2(2C-1)+n^2 = 2n^2C$ operations. 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, 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, we only take operations performed by optical computing into account.

For each pixel combined with spectral filter, assuming that the number of spectral sampling points is C, exposure time is T, then it can perform C multiply operations and $C-1$ additive operations, resulting in the computing speed of $(C+C-1)/T \approx 2C/T$. For an OCL with a spatial resolution of N pixels, the computational speed of the entire OCL can be calculated as follows:

$$s = N \cdot \frac{2C}{T} = \frac{2NC}{T}$$

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. 1). 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.

Under this compression rate, using 9 spectral filters, the sampling points in the spectral dimension is about 216 ($C = 216$). In our implementation, the optical convolutional layer (OCL) has nine convolutional kernels of size 1 and stride 1. For the metasurface-based SCNN, the 3D raw data cube of the natural images has 160×122 superpixels (480 × 366 pixels). The CIS used in our experiment was a Thorlabs CS235MU equipped with a Sony IMX249 sensor. 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. The minimum exposure time is 0.034 m. 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 whole system we have implemented, the cost of data readout and transfer should be considered. 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 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 computing speed of OCL is reduced 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.

For the pigment-based SCNN, it has much higher integration and more spatial pixels. The 3D raw data cube of the natural images has 400×533 superpixels. If we process the raw hyperspectral image of 400×533×216 on electrical computing platform, we need about 176MB storage (stored in 32-bit floating point). 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. The SCNN provides a simple but highly

effective way to sense and process hyperspectral images for various portable terminals. Notably, the pixel size of pigment-based SCNN is only $1.75 \times 1.75\mu m$, resulting in the computing density of about 5.3 TOPS/mm^2. Moreover, the exposure time of the CIS is relatively low (sampling rate is about 37 kHz) compared with high-speed photodetector (sampling rate can even exceed 100 GHz). If we replace the CIS with high-speed PD array, there is still great potential for improvement in computing speed.

Supplementary Note 3. Gradient-based metasurface topology optimization (GMTO) algorithm.

We first adopted freeform-shaped meta-atom metasurfaces[5] to generate millions of different 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: $L(p_1, q_1, \dots, p_N, q_N)$.

Each metasurface can be described by its period $p \in [350,550]$ and shape index $q \in [1,10000]$. Every index q can be mapped to a unique shape $S(q) \in R^{128\times128}$. Considering N metasurfaces $A_{p_k,q_k}, k = 1,2, \dots, N$, each A_{p_k,q_k} has the transmission response $\mathbf{t}_{p_k,q_k} \in R^{M\times1}$. Then, the correlation loss L_{corr} can be calculated as follows:

$$L_{corr} = \max_{i,j=1,2,\dots,N} \{\mathbf{t}_{p_i,q_i}^T \mathbf{t}_{p_j,q_j}\}$$

In addition, we employed a commercial hyperspectral camera to capture 5000 spectra of the positive samples $\mathbf{X}_l \in R^{5000\times M}$ and 5000 spectra of the negative samples $\mathbf{X}_s \in R^{5000\times M}$. Specifically, positive and negative samples represent live and spoof faces in FAS tasks and normal and pathological tissues in the disease diagnosis tasks, respectively. For each A_{p_k,q_k}, we utilized the inter-class variance and intra-class variance to quantify its anti-spoofing ability:

$$d_{p_k,q_k} = \sigma(\mathbf{X}_l \mathbf{t}_{p_k,q_k}) + \sigma(\mathbf{X}_s \mathbf{t}_{p_k,q_k}) - [mean(\mathbf{X}_l \mathbf{t}_{p_k,q_k}) - mean(\mathbf{X}_s \mathbf{t}_{p_k,q_k})]^2$$

where σ denotes the variance and $mean$ denotes the mean value. Then, the FAS loss can be expressed as follows:

$$L_{fas} = \overline{d_{p,q}} = \frac{1}{N}\sum_{k=1}^{N} d_{p_k,q_k}$$

Because the fabrication precision of metasurface nanostructures is limited, we prefer to avoid fabricating two metasurfaces with similar shapes or periods. Although the simulated transmission responses of these two metasurfaces may have a low

correlation, the fabrication error may result in a high similarity between the two actually fabricated metasurfaces. Therefore, we also introduce the fabrication loss into the final loss function to avoid designing similar metasurfaces. The similarity between the two metasurfaces is calculated as follows:

$$sim\left(A_{p_i,q_i}, A_{p_j,q_j}\right) = s(q_i)^T s(q_j) - log\left(1 + 0.005 \times abs(p_i - p_j)\right)$$

where $s(q) \in R^{16384\times1}$ represents the flattened value of $S(q) \in R^{128\times128}$, log represents a logarithm with a base of 10, and abs represents the absolute value. The fabrication loss can be calculated as follows:

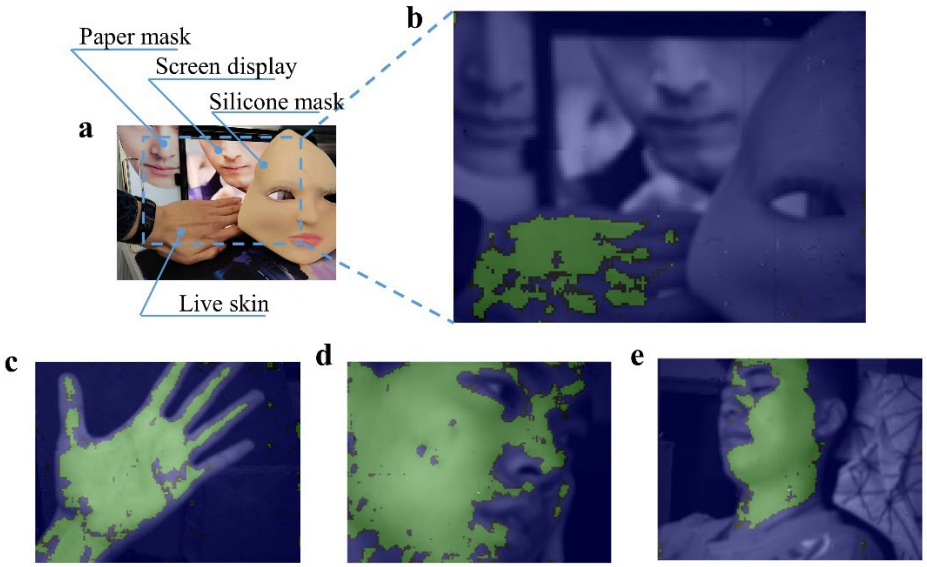
$$L_{fab} = \max_{i,j=1,2,\ldots,N}\{sim\left(A_{p_i,q_i}, A_{p_j,q_j}\right)\}$$

Finally, the total loss was calculated and minimized to obtain an optimized metasurface design.

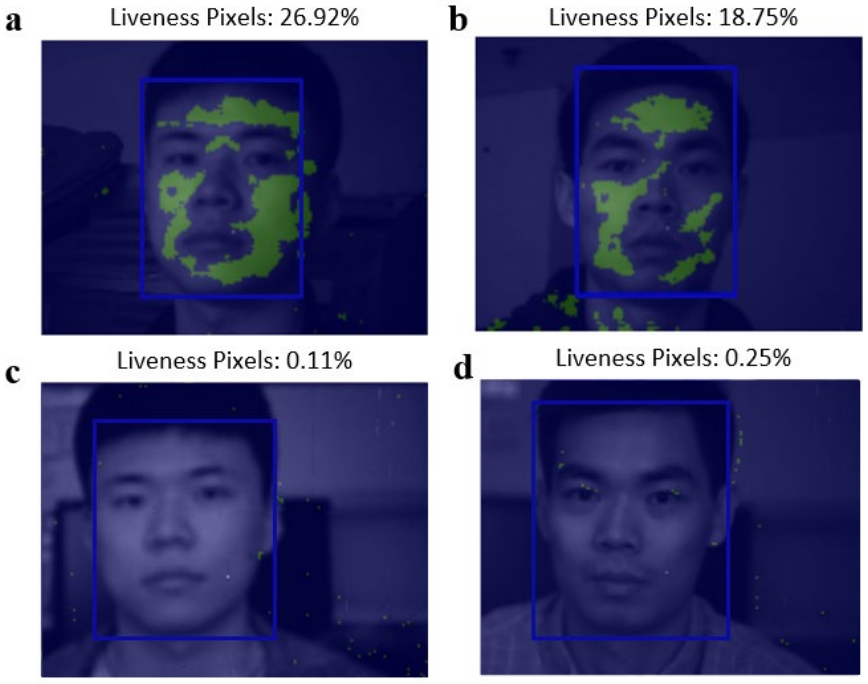
$$L_{total} = \alpha L_{fas} + \beta L_{corr} + \gamma L_{fab}$$

$$\{p_k, q_k\} = arg\min_{\{p_k,q_k\}} L_{total}$$

Supplementary Note 4. Test results for FAS.



Supplementary Fig. 2. Real world test results for pixel-level anti-spoofing liveness detection. a, A test scene consists of a live human hand and several presentation attacks, including paper mask, screen display, and silicone mask. **b,** Predicted results of the test scene by spectral convolutional neural network (SCNN), where the green points represent the live human skin area. **c-d,** Predicted results of other real world scenes.

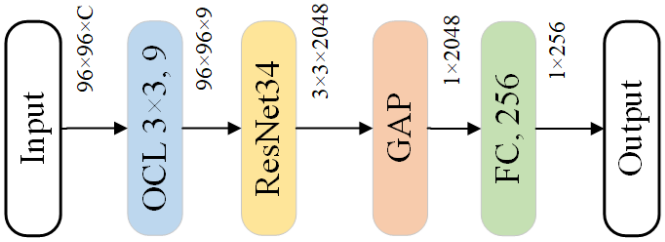


Supplementary Fig. 3. SCNN for image-level FAS on four testing samples. a, b.

The test results of two live faces. **c,** Test result of a face displayed on a screen. **d,** Test results of a printed face.

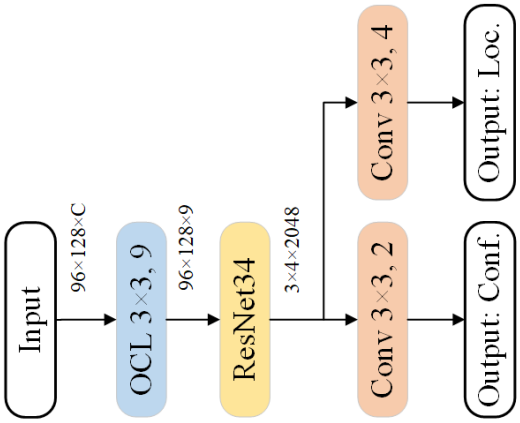
By calculating the liveness area ratio (the proportion of liveness pixels in the detected face bounding box), SCNN can perform reliable image-level FAS. In our experiment results, the ratio of live face image is generally greater than 15% and the ratio of spoof face image is generally less than 1.5%. Therefore, by further processing the pixel-level liveness detection results, this difference in order of magnitude enables the 100% accuracy of FAS. Moreover, by changing and retraining the ENLs, the SCNN can directly achieve 100% accuracy of image-level anti-spoofing.

Supplementary Note 5. Design of ENLs for anti-spoofing face recognition.



Supplementary Fig. 4. Network architecture of the proposed SCNN for face recognition. The OCL layer sensors and simultaneously calculates the convolution results of the hyperspectral facial image. Then ResNet34[6] is adopted to further extract spatial and spectral features from the outputs of OCL layer. Finally, the hyperspectral facial image is embedded into a latent vector $\mathbf{f} \in R^{1\times256}$. We pretrained the network using Arcface[7] loss and stochastic gradient descent optimizer on MS1M[8] dataset. Then we fine-tuned the SCNN on the dataset captured by our own sensor.

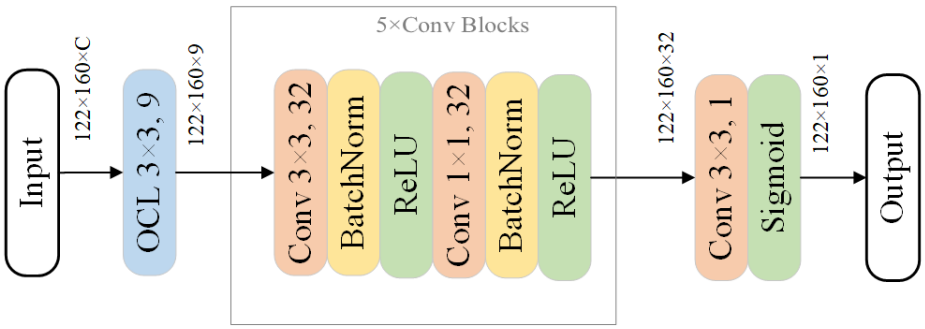
Here, we demonstrate the way to use the SCNN to perform complex vision tasks other than image classification, considering face detection and recognition as examples. In the face recognition task, the OCL inputs were $96 \times 96 \times C$ data cubes. Here C denotes the number of spectral channels and the ENL inputs were $96 \times 96 \times 9$ data cubes. In the face-detection task, the OCL and ENL input sizes were $96 \times 128 \times C$ and $96 \times 128 \times 9$, respectively.



Supplementary Fig. 5. Network architecture of the proposed SCNN for hyperspectral face detection. The function of OCL layer and ResNet34 are the same as described in Supplementary Fig. 4. However, on the outputted feature maps of ResNet34, 2 convolution kernels of size 3×3 are used to calculate the confidence

(Conf.) of predicted label. Another four convolution kernels of size 3×3 are used to predict the location (Loc.) of bounding boxes. SSD[9] loss was adopted to train the network. The network was first trained on Wider Face[10] dataset and then fine-tuned on the dataset captured by our sensor. These works show that by simply changing the network layers implemented on CPU/GPU, SCNN can perform various advance computer vision tasks on hyperspectral image.

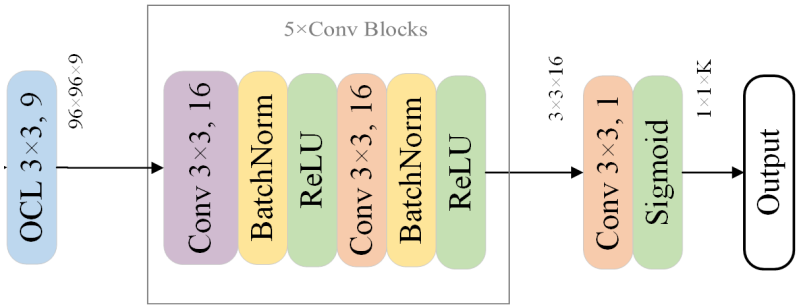
Supplementary Note 6. Design and training details of the ENLs for metafurface-based SCNN.



Supplementary Fig. 6. Network architecture of our SCNN for pixel-level disease detection and liveness detection.

The inputs of SCNN are the 3D raw data cube of natural images. The OCL layer sensors and calculates the results of convolution at the same time. It has nine convolutional kernels of size three and stride nine. The feature maps outputted by OCL are further processed by the following electrical layers on CPU. The final outputs of $122 \times 160 \times 1$ represents the pixel-level detection results.

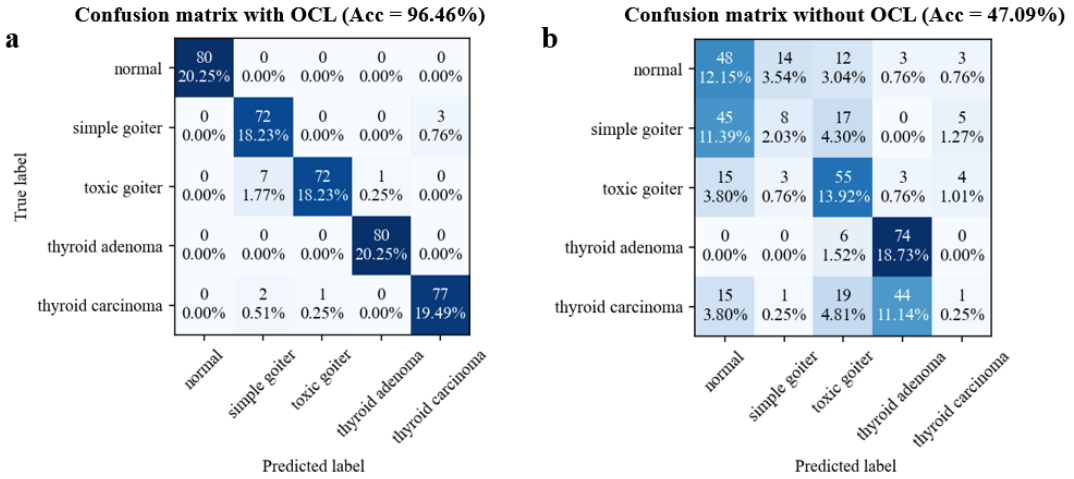
To train the ENLs for disease detection, we employed our SCNN chip to capture 2990 samples of thyroid histological sections through a microscope and randomly selected 250 samples to serve as the test set, using the rest as the training set. To train the ENLs for FAS, we employed our SCNN chip to capture more than 200 samples of live skin and spoof material in the real world. We then obtained pixel-level annotations of the feature cubes by manual labeling. Pixels located on live human skin were labeled positive, whereas non-live pixels on various materials, including environmental objects and spoof materials such as silicone masks, latex masks, and resin masks, were labeled negative. Finally, ENLs were trained on the labeled dataset. Then we employed our SCNN chip to capture images, obtained a test set containing 108 test samples from 31 individuals, and evaluated the performance of the SCNN framework.



Supplementary Fig. 7. Network architecture of our SCNN for image-level disease diagnosis and FAS.

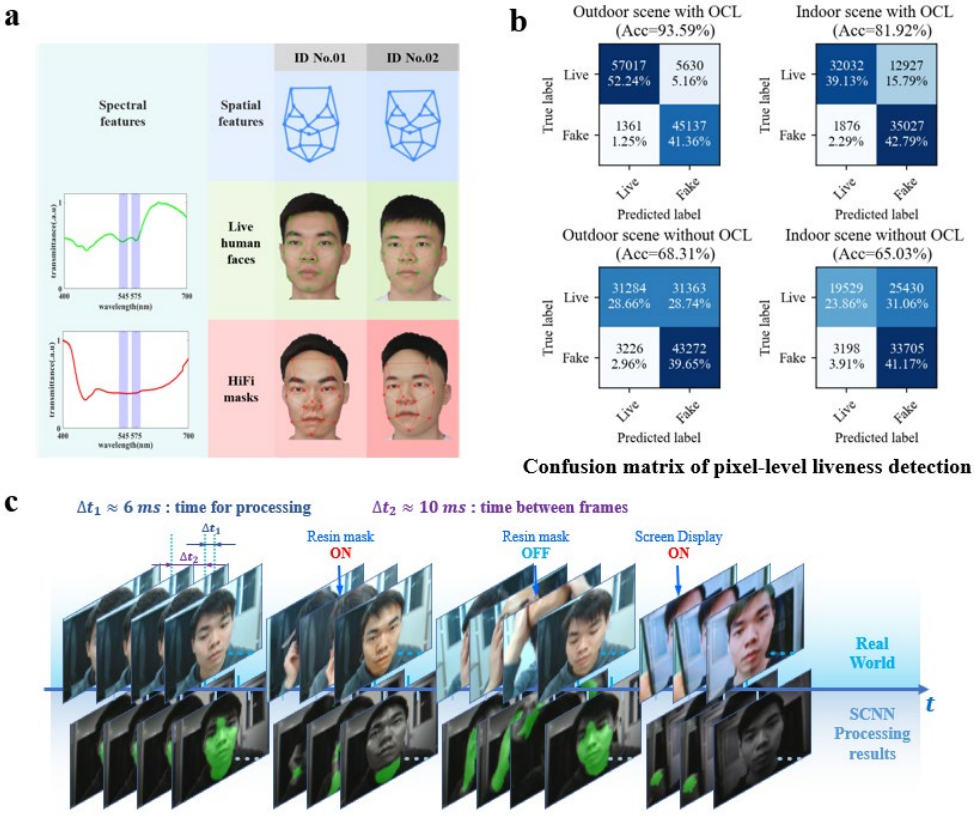
The inputs of SCNN are 3D raw data cube of natural images with size $96 \times 96 \times C$, where C denotes the number of spectral channels. The OCL has nine convolutional kernels of size three and stride nine. Therefore, the feature maps outputted by OCL has the size $96 \times 96 \times 9$. In the ENLs, each convolutional block contains two convolutional layers with 16 kernels of size three, two BatchNorm layers, and two ReLU layers. The strides of the two convolutional layers are one and two. The final outputs of $1 \times 1 \times K$ represents the pixel-level detection results. Here, K represents the number of classes, which is five in thyroid disease diagnosis task and two in FAS task.

Supplementary Note 7. Experimental results of pigment-based SCNN chip.



Supplementary Fig. 8 Experimental results of thyroid histological section diagnosis by the SCNN. a, Our SCNN achieves an accuracy of 96.46% on the thyroid disease diagnosis task. **b,** The classification accuracy is only 47.09% without the OCL.

For FAS, 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. The confusion matrix of the classification results on the testing dataset is shown in Supplementary Fig. 8a. Note that all of the misclassified samples are between the four diseases, which indicates that the four thyroid diseases are not completely independent and that there may be complicating pathologies. Moreover, if we distinguish only normal samples from diseased samples, the accuracy on the testing dataset is 100%. Furthermore, we 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%. This indicates that spectral information is vital to diagnosis. Our OCL enables powerful spectral sensing capabilities, and the features acquired by OCL are effective for the precise diagnosis of pathological sections. Moreover, the diagnosis process does not require a microscope.



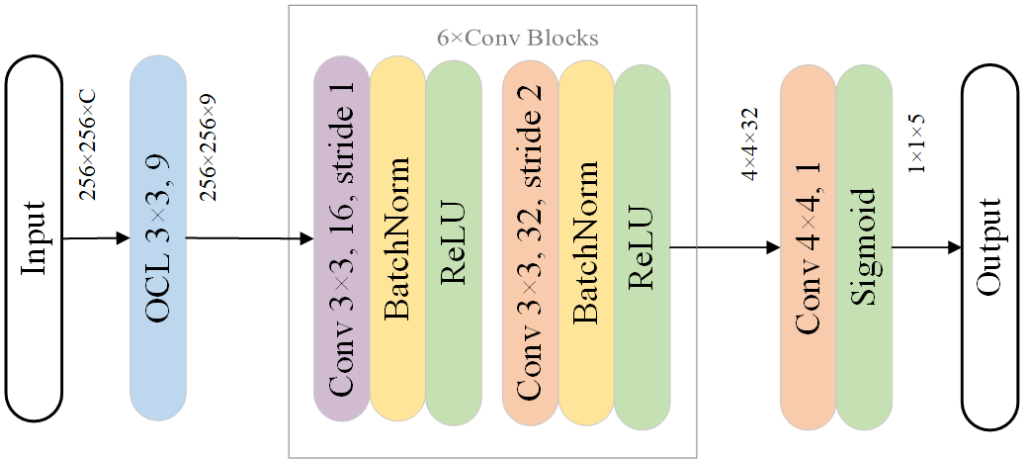
Supplementary Fig. 9 SCNN chip can be used for pixel-level anti-spoofing liveness detection. a, Our SCNN chip can combine spectral features with spatial features and perform reliable anti-spoofing face recognition. **b,** Confusion matrix for the pixel-level liveness detection results with and without OCL under sunlight and white LED light. **c,** Video-rate liveness detection based on the SCNN chip can detect high-fidelity lifelike masks effectively. The detected pixels of live skin are marked in green. More results can be found in the Supplementary Video 1.

In addition to histological section diagnosis, we employed the proposed SCNN for FAS to further study its capability for computer vision tasks. 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, discriminative features can be extracted to detect HiFi masks when powered by our MMI. The broadband incoherent natural light that includes two spatial dimensions and one spectral dimension that are first captured and processed by the OCL. The size of the feature maps output by OCL is

$400 \times 533 \times 9$. Then, the feature maps are further processed by several ENLs, and we can obtain the anti-spoofing pixel-level liveness detection results. Our SCNN chip can combine spectral features with spatial features and perform reliable anti-spoofing face recognition (Supplementary Fig. 9a). To test its real-world performance, we collect data samples in both outdoor scene and indoor scene. Supplementary Fig. 9b shows the confusion matrix for the test samples. The SCNN achieves accuracies of 93.59% and 81.92% in outdoor and indoor scenes, respectively. In the outdoor scene, the ambient light source is sunlight. Because sunlight covers a wider spectral band, it also has a high intensity in the near-infrared band, which is useful for anti-spoofing and leads to better results. The indoor scene is mainly illuminated by artificial light sources such as LED. They usually have weak intensity in the near-infrared band. To improve the performance in the indoor scene, we can adopt a wide-band lamp as the fill light. And we will also further optimize the hardware design specifically for indoor scenes. If we remove the OCL, the accuracies decrease to 68.31% and 65.03%, respectively, which demonstrates that the spectral information obtained by OCL is of vital importance.

Furthermore, we can conduct image-level FAS based on the pixel-level liveness detection results. By applying an additional face detection procedure, we can get the bounding boxes of the faces and then calculate the averaged value of the pixel-level liveness detection results for each bounding box. As for the live and spoof faces shown in Fig. 4g, the averaged values are 0.6052 and 0.0016 respectively. In this way, we can get almost 100% image-level anti-spoofing accuracy. To show the real-world FAS capabilities, we employed the designed SCNN chip to perform real-time anti-spoofing pixel-level liveness detection at different video frames (Supplementary Fig. 9c). The frame rate of the results is almost only limited by the CIS exposure time. The HiFi masks can be easily detected at the pixel level (more results can be found in the Supplementary Video 1). Thus, the proposed SCNN framework is expected to be widely used in real-world MMI applications. The results indicate that by simply redesigning and retraining the ENLs according to the needs of specific tasks, the function of the SCNN can be customized as the disease diagnosis task and the liveness

detection task performing at image and pixel levels. 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. It can combine the advantages of optical and electrical computing.

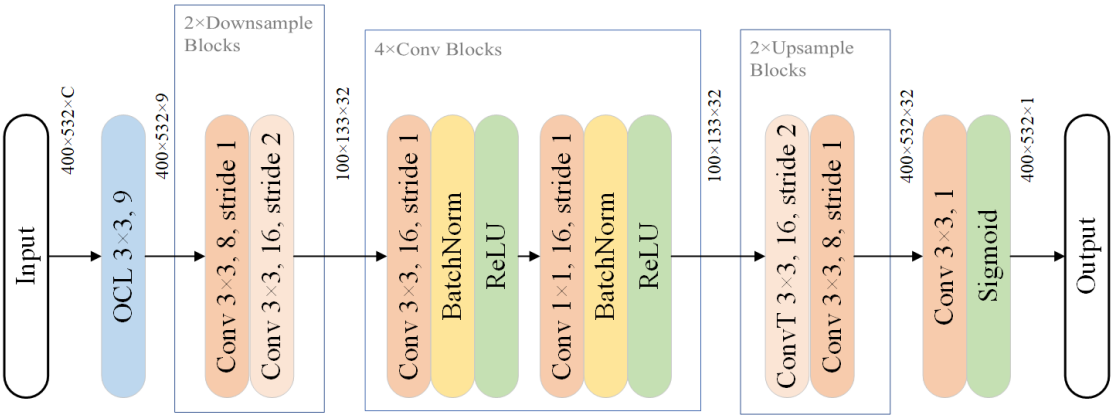


Supplementary Fig. 10. Network architecture of our SCNN for image-level disease diagnosis.

The inputs of SCNN are 3D raw data cube of natural images with size $256 \times 256 \times C$, where C denotes the number of spectral channels. The OCL has 9 convolutional kernels of size 1×1 and stride 1. Therefore, the feature maps outputted by OCL has the size $256 \times 256 \times 9$. In the electrical network layers (ENLs), each convolutional block contains two convolutional layers. Each convolutional layer is followed by a BatchNorm layer and a ReLU layers. The first convolutional layer has 16 kernels of size 3×3 and stride 1. The second convolutional layer has 32 kernels of size 3×3 and stride 2 to perform downsampling. The final outputs of $1 \times 1 \times 5$ represents the 5-class classification results.

We utilize 100 histological sections from 100 different patients to collect the dataset. The 100 sections contain 5 categories (normal, simple goiter, toxic goiter, thyroid adenoma, and thyroid carcinoma), each with 20 sections. We collected about 500 samples from these 100 different sections. 80 sections were employed as training set and the remaining 20 sections were testing set. Then we employ our SCNN sensor

to capture the 9-channel feature maps of these sections and build the training and testing set. Each section has been sampled several times. The raw outputs of the SCNN sensor have the size of $400 \times 533 \times 9$. We randomly crop the raw outputs to the size of $256 \times 256 \times 9$ to perform data augmentation. Finally, the ENLs are trained using the collected dataset. The Adam optimizer and cross-entropy loss are adopted to train the network and the learning rate is 0.001.



Supplementary Fig. 11. Network architecture of our SCNN for pixel-level liveness detection.

The feature maps outputted by OCL are further processed by the following electrical layers on CPU/GPU. After several convolutional layers, the final outputs have size $400 \times 532 \times 1$ and represent the pixel-level anti-spoofing liveness detection results. To train the ENLs for FAS, we employ our SCNN sensor to capture more than 200 samples of live skin and spoof material in the real world. The spoofing materials include silicone masks, paper masks, resin masks, and screen display. We then obtain pixel-level annotations of the feature cubes by manual labeling. Pixels located on live human skin are labeled as positives, whereas non-live pixels on various materials, including environmental objects and spoof materials are labeled as negatives. Finally, ENLs are trained on the labeled dataset. The ENLs can achieve a processing speed of about 14 frames per second (fps) at a batch size of 1, or about 20 fps at a batch size of 8, on an Intel Core i7-11700 @2.5GHz CPU. As the computing of OCL and ENLs can

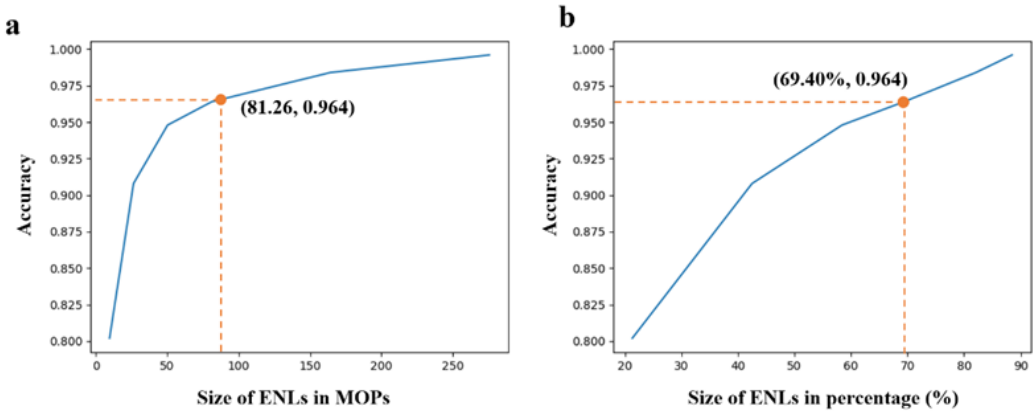
be asynchronous. That is, while the ENLs are processing the current frame, the OCL is capturing the next frame. In this way, the real-world performance of SCNN can achieve video rate.

Supplementary Note 8. Additional analysis on disease diagnosis application of the pigment-based SCNN.

The disease diagnosis is a 5-class classification task. Assuming that the confusion matrix $\mathbf{M} = \{m_{ij}\} \in R^{5\times5}$, each row of $\mathbf{M}$ represents a true label and each column of $\mathbf{M}$ 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 $\frac{\mathrm{tr}(\mathbf{M})}{\mathrm{sum}(\mathbf{M})}$, 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.

	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%

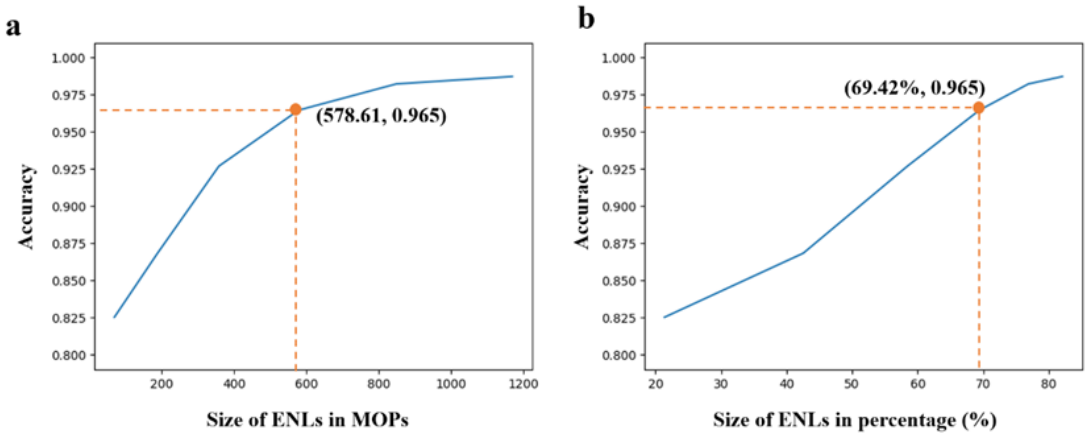
Supplementary Note 9. The relationship between the size of ENLs and the performance of the whole SCNN.



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. 12. 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 of the computer vision tasks.

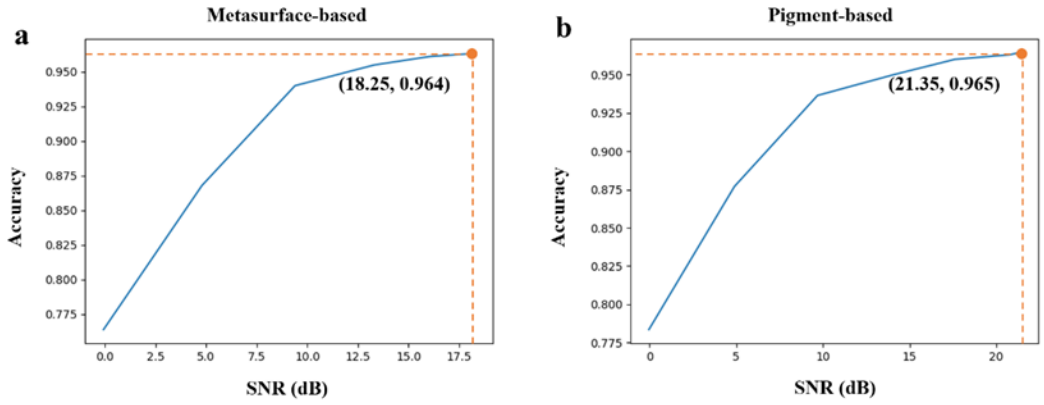
These results indicate that, although larger ENLs can lead to better results, we can still reach a considerable performance using small-size ENLs because the OCL can provide powerful in-sensor feature extracting capabilities. Similar results can also be achieved by the pigment-based SCNN, as shown in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

Supplementary Note 10. Study on the signal-to-noise (SNR) of the fabricated SCNN chip.

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.

References

[1] Xiong, J. *et al*. Dynamic brain spectrum acquired by a real-time ultraspectral imaging chip with reconfigurable metasurfaces. *Optica* **9**, 461-468 (2022).

[2] 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).

[3] Yang, J. et al. Deep-learning based on-chip rapid spectral imaging with high spatial resolution. *Chip* **2**, 100045 (2023).

[4] Yako, M., Yamaoka, Y., Kiyohara, T. et al. Video-rate hyperspectral camera based on a CMOS-compatible random array of Fabry–Pérot filters. *Nat. Photon*. **17**, 218–223 (2023).

[5] Yang, J. *et al*. Ultraspectral imaging based on metasurfaces with freeform shaped meta-atoms. *Laser & Photonics Reviews* **16**, 2100663 (2022).

[6] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 770-778, (2016).

[7] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690-4699, (2019).

[8] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, Springer, 87-102 (2016).

[9] Liu, W. et al. SSD: Single shot multibox detector. In *European conference on computer vision*. Springer, Cham, 21-37, 2016.

[10] Yang, S., Luo, P., Loy C., and Tang, X. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, (2016).

[11] Xu, Y., Lu, L., Saragadam, V. *et al.* A compressive hyperspectral video imaging system using a single-pixel detector. *Nat Commun* **15**, 1456 (2024). https://doi.org/10.1038/s41467-024-45856-1