# Fourier-space Diffractive Deep Neural Network: Supplementary Material

Tao Yan[1,*], Jiamin Wu[1,*], Tiankuang Zhou[1,2,*], Hao Xie[1], Feng Xu[3], Jingtao Fan[1], Lu Fang[2], Xing Lin[1,4,†], and Qionghai Dai[1,‡]

[1]*Department of Automation, Tsinghua University, Beijing, 100084, P. R. China*

[2]*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, P. R. China*

[3]*School of Software, Tsinghua University, Beijing 100084, P. R. China*

[4]*Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA*

[*]*These authors contributed equally to this Letter.*

*E-mail: [†]linx2019@tsinghua.edu.cn, [‡]qhdai@tsinghua.edu.cn*

## Supplementary Methods

**Training Fourier-space D$^2$NN:** We propose to train a diffractive deep neural network (D$^2$NN) in Fourier-space since every different object will have a very different Fourier transform pattern. The dual 2*f* system is used for performing Fourier transform and inverse Fourier transform, the parameters of which determine the optical magnification, modulation resolution, and the size of a network. Both the saliency detection network and classification network were trained under a visible wavelength ($\lambda$=532 nm), which offers the network with a high resolution and integration property. For the task of saliency detection trained with Cell dataset at unit magnification in the main text (Fig. 2), the N.A. and focal length of the first 2*f* system was set to be the same as the second 2*f* system: N.A.$_1$=N.A.$_2$=0.112, $f_1$= $f_2$=10 mm. Considering the physical fabrication and N.A. matching, the neuron size of the network was set to be 2 μm to have sufficient sampling rate for the modulation layer. The five-layer phase modulation network for saliency detection was trained by packing 800×800 neurons over each layer of the network, covering an area of 1.6 mm×1.6 mm per layer. During the preparation of Cell dataset, the pathology slide image of cell type 1 [39], which has the pixel number of 36001×30261, was divided into 3500 patches with 200×200 pixels for each patch and upsampled two times with

boundary padding for matching the size of a network. We randomly selected 2750 patches as the training dataset, 250 patches as the validation dataset, and 500 patches as the testing dataset. All cell images were converted into the grayscale due to the using of a single wavelength. For the training of saliency detection network with CIFAR-10 dataset [43], we changed the focal length of the dual 2*f* system to 2mm and neuron number of each layer to 160×160 (layer size of 0.32 mm×0.32 mm) while keeping other network settings the same. Specifically, we used the CIFAR-10 dataset "Cat" category (5000 images, each with a size of 32×32) to train the network, where images were converted into grayscale and resized to $100 \times 100$ with the boundary padding to match the network size. We used the online video sequences [42] and CIFAR-10 dataset "Horse" category as the testing images. Similarly, for the task of classification in the main text, we set N.A.$_1$=N.A.$_2$=0.16, $f_1$= $f_2$=1 or 4 mm, with the neuron size of 1 μm. The number of neurons on each layer of a phase modulation network was set to be 200×200 with zero paddings to match the diameter of the 2*f* system. The classification network was trained with the MNIST (Modified National Institute of Standards and Technology) handwritten digit (0, 1, …, 9) dataset [18], which has 55,000 training images, 5000 validation images, and 10,000 testing images with each image size of 28×28 pixels. The image size of the MNIST dataset is upsampled three times with boundary padding to match the network size. The successive layer distance of both saliency detection and classification Fourier-space $D^2$NNs was optimized and set to be 100 μm.

Similar to [18], the proposed framework was implemented using TensorFlow machine learning library (Google Inc.). Both the phase and amplitude of the modulation layer in principle can be the learnable parameters for each neuron on each layer. In this work, we used the phase-only modulation due to its simplicity of the physical implementation in contrast to the complex modulation, where each layer can be approximated as a thin optical element, and the distance between successive layers decides the receptive field of neurons. The free-space propagation is implemented with the angular spectrum method. To facilitate the physical fabrication of F-$D^2$NN with direct laser writing, a sigmoid function was used to constrain the phase coefficient of each neuron within a certain range, i.e., $0\text{-}2\pi$ and $0\text{-}\pi$, for saliency detection and classification networks, respectively. The ground truth for training saliency detection network

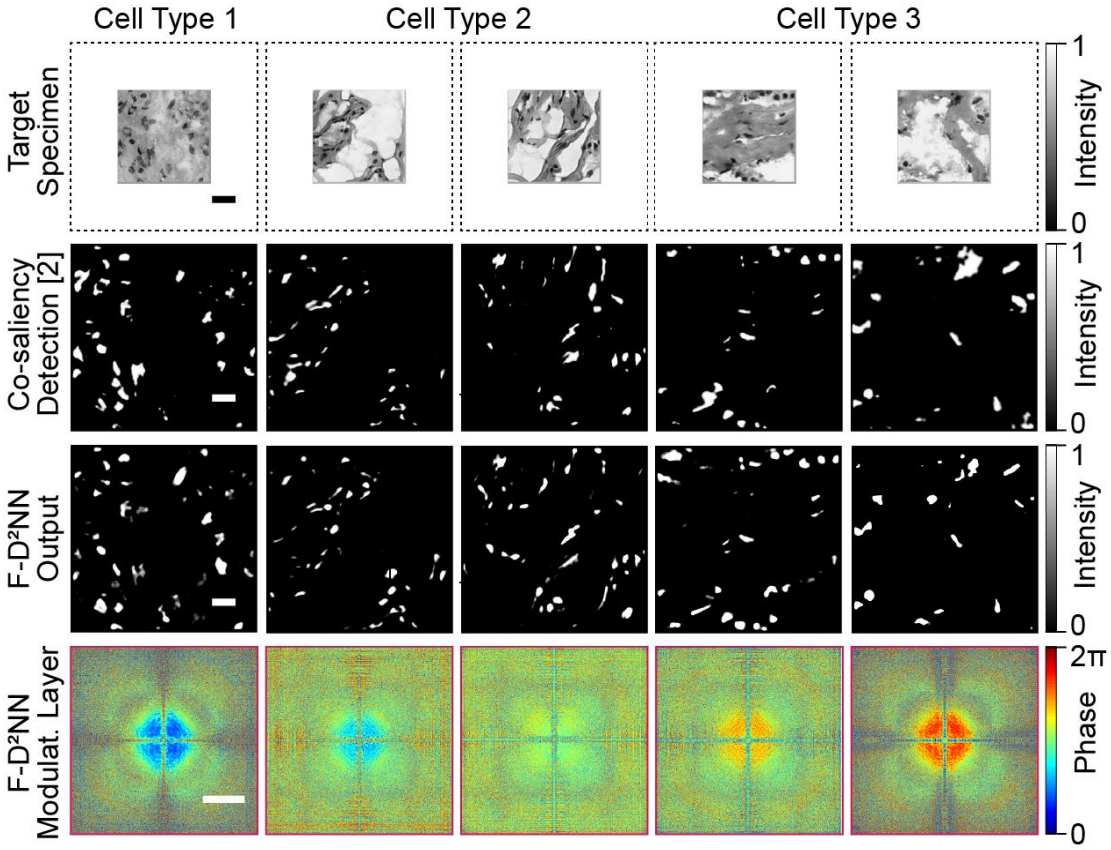
was obtained by using the state-of-the-art algorithms, including co-saliency detection algorithm [27] and robust background detection (RBD) algorithm [30]. The ground truth for classification network was obtained by setting ten detectors regions (corresponding to ten digits) with each detector width of 12 μm. The stochastic gradient descent algorithm (Adam optimizer) [1] is used to back-propagate the errors and update the diffractive layers of the network to minimize the loss function. With a learning rate of 0.01 and batch size of 10, the training iteration converged after 100 epochs and 30 epochs for saliency detection network and classification network, respectively. We implemented the networks using Python version 3.6.0. and TensorFlow framework version 1.11.0 (Google Inc.). Using a desktop computer (Nvidia TITAN XP Graphical Processing Unit, GPU, and AMD Ryzen Threadripper 2990WX CPU with 32 Cores, 128GB of RAM, running with a Microsoft Windows 10 operating system), it took approximately 20 hours to train the saliency detection and classification networks with the above-outlined implementation. During the numerical testing, it took 1 minute to inference 100 testing images for Cell dataset and 1 minute to inference MNIST testing dataset.

**Optical Nonlinearity Property:** The proposed F-D$^2$NN uses coherent illumination as an energy source to generate the complex optical field propagation and perform complex-value optical computing. The complex activation function used in this work creates the nonlinear phase modulation with respect to the intensity variations, which is achieved by employing a photorefractive crystal (SBN:60) nonlinear material [35, 36]. The photorefractive effect can lead to this type of optical nonlinear response much stronger than other optical nonlinear effects, such as the Kerr effect. The required incident light intensity for SBN crystal to turn on the nonlinearity effect is only at the scale of ~0.1 mW/mm$^2$, and the response time of the SBN crystal depends on the incident light intensity. Besides, in contrast to the amplitude or intensity modulations, using phase modulation doesn't consume energy. The origin of the nonlinearity is the spatial variant optical intensity, which generates free charge carriers through photoionization. Free charge carriers then change the local electric field distribution, and the refractive index is finally shifted accordingly that results in phase modulation [37]. The index change of SBN can be modeled as: $\Delta n = \kappa E_{app} < I > /(1+< I >)$, where $< I >$ is the intensity perturbation above a spatially homogeneous background intensity $< I_0 >$, $E_{app}$ is

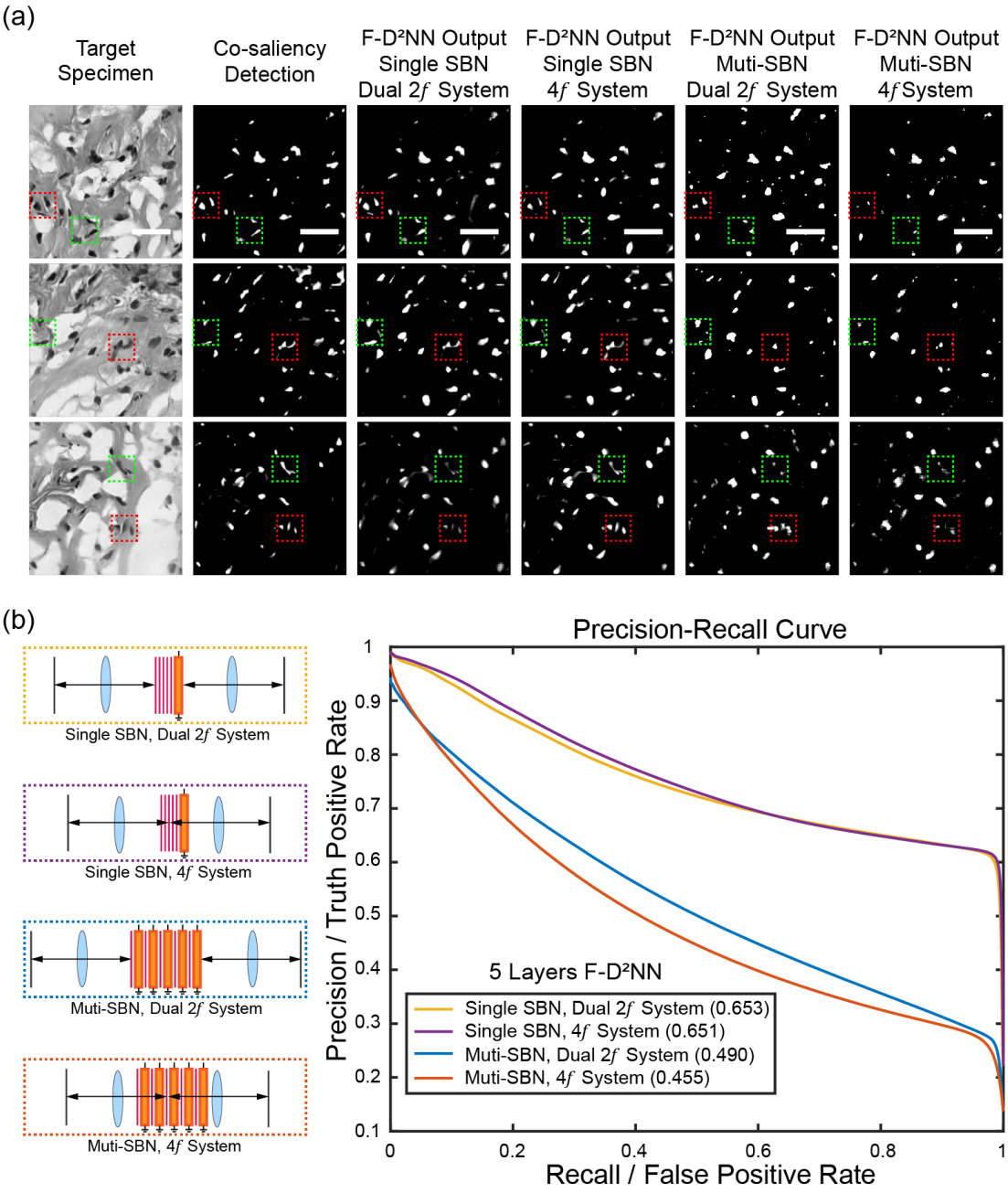
the applied electric field along c axis, and $\kappa = n_0 r_{33}(1+< I_0 >)$ is a constant depending on the base index of refraction $n_0$, the electro-optic coefficient $r_{33}$, and $< I_0 >$. In practical, the crystal has voltage-dependent nonlinearity. To generate enough strength of nonlinearity and consider the implementation feasibility, we set the thickness of photorefractive crystal to be 1 mm and voltage on a crystal to be 972 V, with which the phase variation of nonlinear material is between 0~$\pi$ and can be formulated as $\Delta\emptyset(I) = \pi < I >/(1+< I >)$.

Although the nonlinear phase modulation property of the crystal with respect to the intensity variation doesn't consume energy, we estimate the energy loss of photorefractive crystal from three main aspects: absorption, reflection, and dark current. (1) Absorption. The absorption coefficient $\alpha$ of nominally undoped SBN:60 is $\sim 0.1\ \mathrm{cm}^{-1}$ at the wavelength of 532 nm [37], and the thickness of SBN:60 is $\sim 1$ mm in our experimental setup. Therefore, the transmission rate can be modeled as $\mathrm{T} = e^{-\alpha d} \sim 0.99$, which means ~1% of optical energy will be absorbed. (2) Reflection. When the light is incident perpendicular to the crystal surface, the reflection coefficient caused by the refractive index difference of the air-crystal interface can be calculated as $((\mathrm{n}_1 - \mathrm{n}_2)/(\mathrm{n}_1 + \mathrm{n}_2))^2 \approx 0.16$, where $\mathrm{n}_1 = 1.00$ represents the refractive index of air, and $\mathrm{n}_2 \approx 2.33$ represents the refractive index of SBN:60. In practical, such reflection can be reduced to ~1% by an anti-reflective coating. (3) Dark Current. The dark conductivity of SBN:60 $\sigma_d$ is $\sim 10^{-10}\ \Omega^{-1}\mathrm{cm}^{-1}$ [37]. Supposing that the crystal size is $\mathrm{a} \times \mathrm{b} \times \mathrm{c} = 2\ \mathrm{mm} \times 2\ \mathrm{mm} \times 1\ \mathrm{mm}$ (2mm along c axis), the resistance along the electric field of the crystal is $R = \mathrm{b}/\sigma_d \mathrm{ac} \sim 10^{11}\ \Omega$. The voltage along c axis $U$ is $\sim 1$ kV, so the electric power is $P_E = U^2/R \sim 0.01$ mW. Therefore, if the illumination power is ~10 mW, the energy loss by the dark current can be ignored compared with the amount of absorption and reflection (2% of the total power), and the total energy loss of the photorefractive crystal can be estimated as ~0.2mW.

## Supplementary Figures

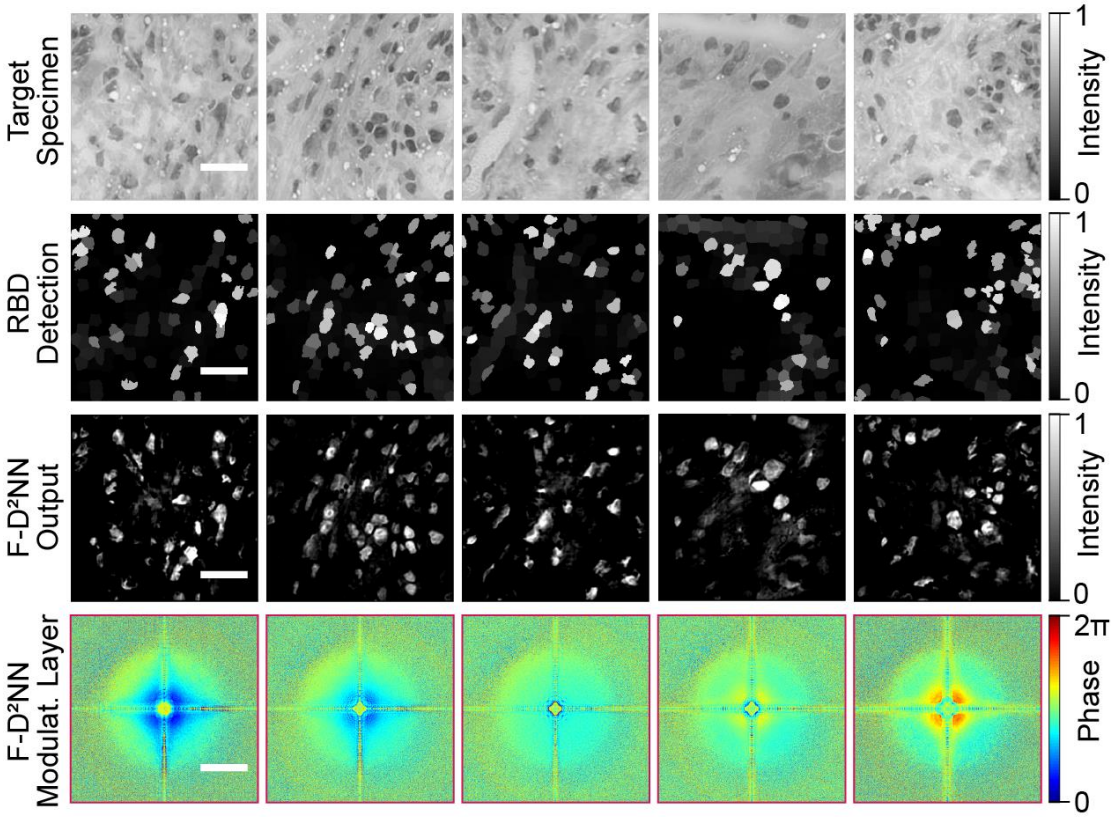



Supplementary Figure S1: All-optical saliency detection with Fourier-space D$^2$NN on a 2× magnification optical system. We demonstrate the success of Fourier-space D$^2$NN for cell segmentation by training it to perform a saliency detection mapping function. During the implementation, the focal length of the second $2f$ system was set to be two times of the first $2f$ system: $f_2 = 2f_1 = 20$ mm; other design parameters and datasets were set to be the same as Fig. 2 of the main text. By simply placing the trained Fourier-space D$^2$NN on the Fourier plane of the designed 2× magnification imaging system, the optical system immediately performs an all-optical saliency detection of the specimen, which further demonstrates the potential application of the proposed method in microscopic imaging. The maximum F-measure of the average precision-recall curves on the result of a testing dataset is 0.464 (Scale bar: 200μm for the first to the third row, 400 μm for the fourth row. Cell scales are optically magnified.)

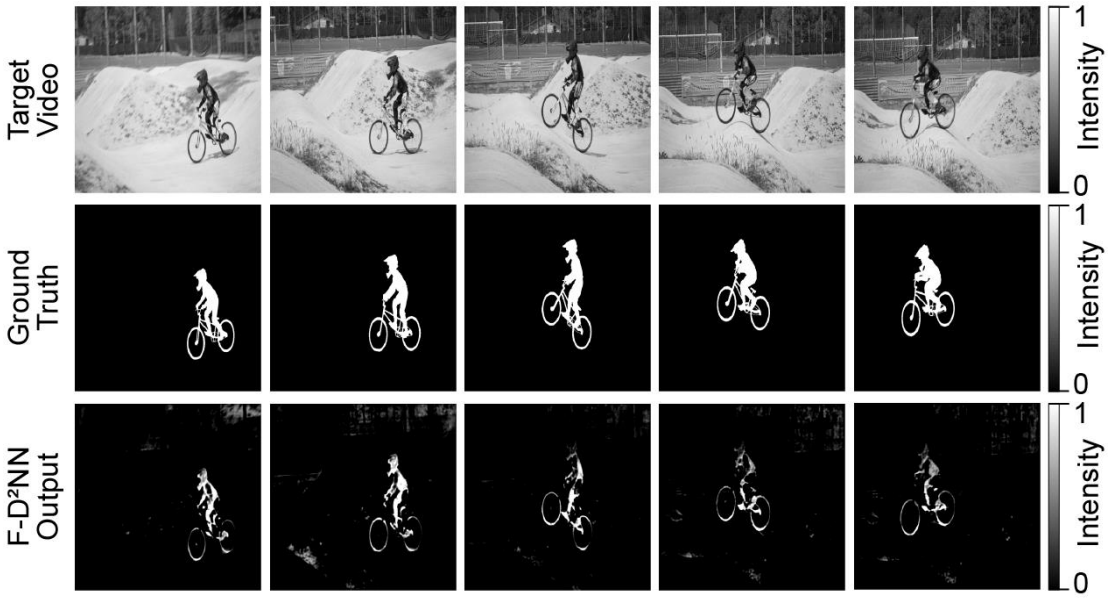



Supplementary Figure S2: The performance of five-layer F-D$^2$NN for saliency detection on the dual 2*f* system and 4*f* system configured with a single SBN nonlinear layer and multiple SBN nonlinear layers. In dual 2*f* system setting, the Fourier planes of the first and second 2*f* system are set to be on the first and last layer of F-D$^2$NN, respectively; on the other hand, the Fourier planes of two 2*f* systems can be superposed as a 4*f* system and set to be in the middle of the F-D$^2$NN. The multi-SBN configuration is implemented by incorporating the SBN to individual modulation layer, while the single SBN configuration simplifies the design by placing a single SBN at the end of modulation layers. The corresponding four different types of configurations

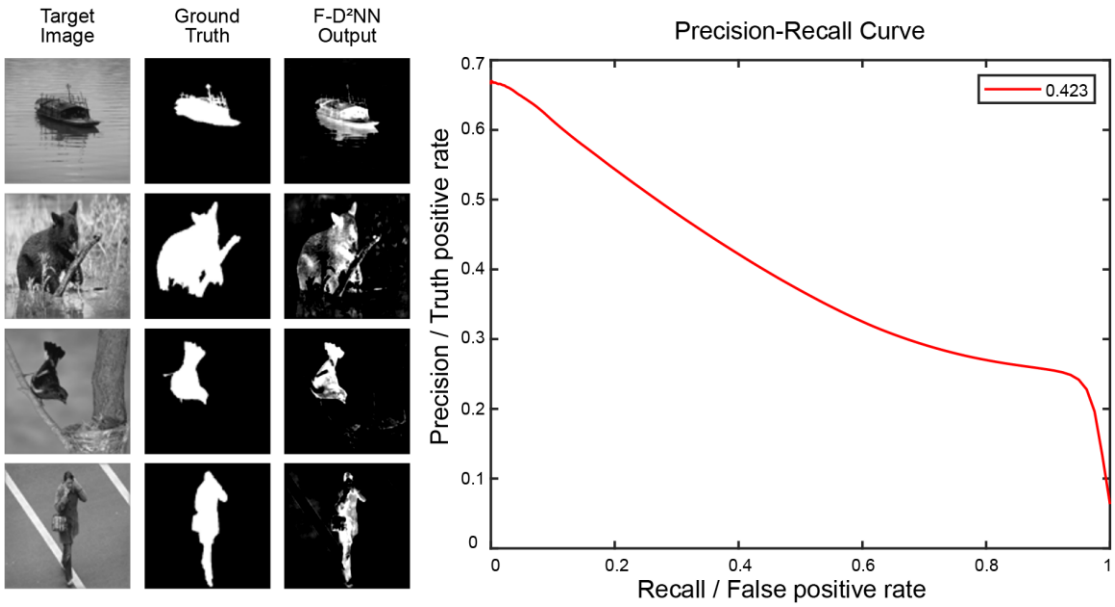
are demonstrated in the left half of (b). The five-layer F-D$^2$NN under four different configurations are trained on the same pathology slide datasets (Cell Type 1) we used in the main text for the task of salient object detection, the corresponding average precision-recall curves on the testing datasets are shown in the right half of (b). The experiment results show the comparable performance of the dual 2*f* system and 4*f* system on 5-layer F-D$^2$NN. On the other hand, because the significantly increasing of layer distance decreases the effective diffractive modulation resolution, and the using of multi-SBN nonlinear layer introduces more optical aberrations, the maximum F-measures of single SBN configurations are 0.653 and 0.651 for dual 2*f* system and 4*f* system, respectively, which significantly outperforms the multi-SBN configurations (0.490 and 0.455 for dual 2*f* system and 4*f* system, respectively). The representative saliency detection results are shown in (a); by using the same testing targets (Cell Type 2, first column), the F-D$^2$NN results on four different configurations are shown in the last four columns. We adopted the dual 2*f* system with a single SBN configuration in this work, which achieves comparable saliency detection and cell segmentation result (third column) with respect to the co-saliency detection algorithm (second column). (Scale bar: 200 μm for the first to the fourth row. Cell scales are optically magnified.)



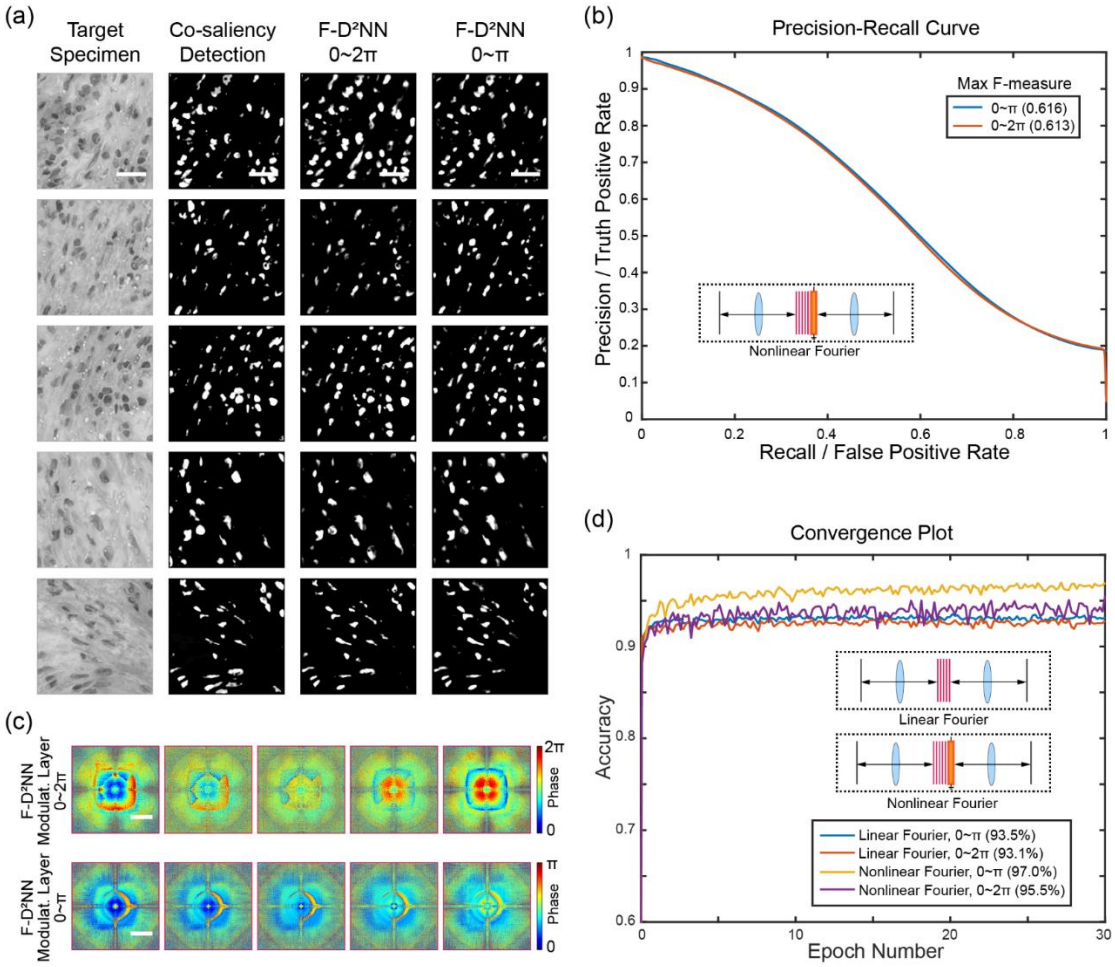

Supplementary Figure S3: Cell segmentation with Fourier-space $D^2NN$ by training it with a different ground truth saliency detection method. In this experiment, we applied the robust background detection (RBD) algorithm [29] to generate ground truths for the input pathology slide images, which were used to train the network for performing the task of salient object detection. The network settings were kept the same as Figure 2 in the main text. The RBD algorithm created various false positive regions and sharp boundary for the detected areas due to the using of boundary connectivity measurement, which placed constraints on the performance of the Fourier-space $D^2NN$ could achieve. We validated the trained network (last row) by testing the images of Cell Type 1 from the testing dataset (first row) and found that the proposed optical neural network was able to reject the false positive regions and generated more meaningful cell detection results (third row) compared with the RBD detection results (second row). The smoother boundary of results is created by the limited numerical aperture of the network as an optical imaging system. The maximum F-measure of an average precision-recall curve over the testing dataset is 0.377. Notice that Fourier-space $D^2NN$ was implemented under single wavelength at the visible spectrum in this work; thus, both RBD and optical saliency detection methods were using grayscale images for fair comparisons. (Scale bar: 200 μm for the first to the third row, 400 μm for the fourth row. Cell scales are optically magnified.)




Supplementary Figure S4: Testing the Fourier-space $D^2NN$ on the video object segmentation benchmark dataset, i.e., DAVIS (Densely Annotated Video Segmentation) [38]. To demonstrate the performance of the proposed approach for saliency detecting and segmenting of high-speed moving targets, we converted the " bmx-bumps" video from DAVIS 2016 dataset into a grayscale video and used it as an input to the trained Fourier-space $D^2NN$ of the main text for salient object detection frame by frame. In this example, we tested the first 30 frames of the " bmx-bumps" video sequence as shown in the Supplementary Video 2. The optical neural network successfully segmented out the foreground moving target (last row) for the input video sequence (first row); and after the sensor measurement, all-optical saliency detection result can be further used for tracking, recognition or other computer vision tasks using an electronic computer. We calculated the average precision-recall curve for Fourier-space $D^2NN$ outputs with respect to the ground truth segmentation results provided in the dataset (second row) over the tested frames, the maximum F-measure of which is found to be 0.498. The maximum F-measure of the state-of-the-art convolutional neural network for saliency detection [27] on the DAVIS dataset with color information is about 0.7. The numerical experimental results demonstrate the application of the proposed framework for low power consumption, light-speed detecting, and segmenting of saliency targets that can potentially revolutionize the imaging and computing pipeline in various computer vision tasks.

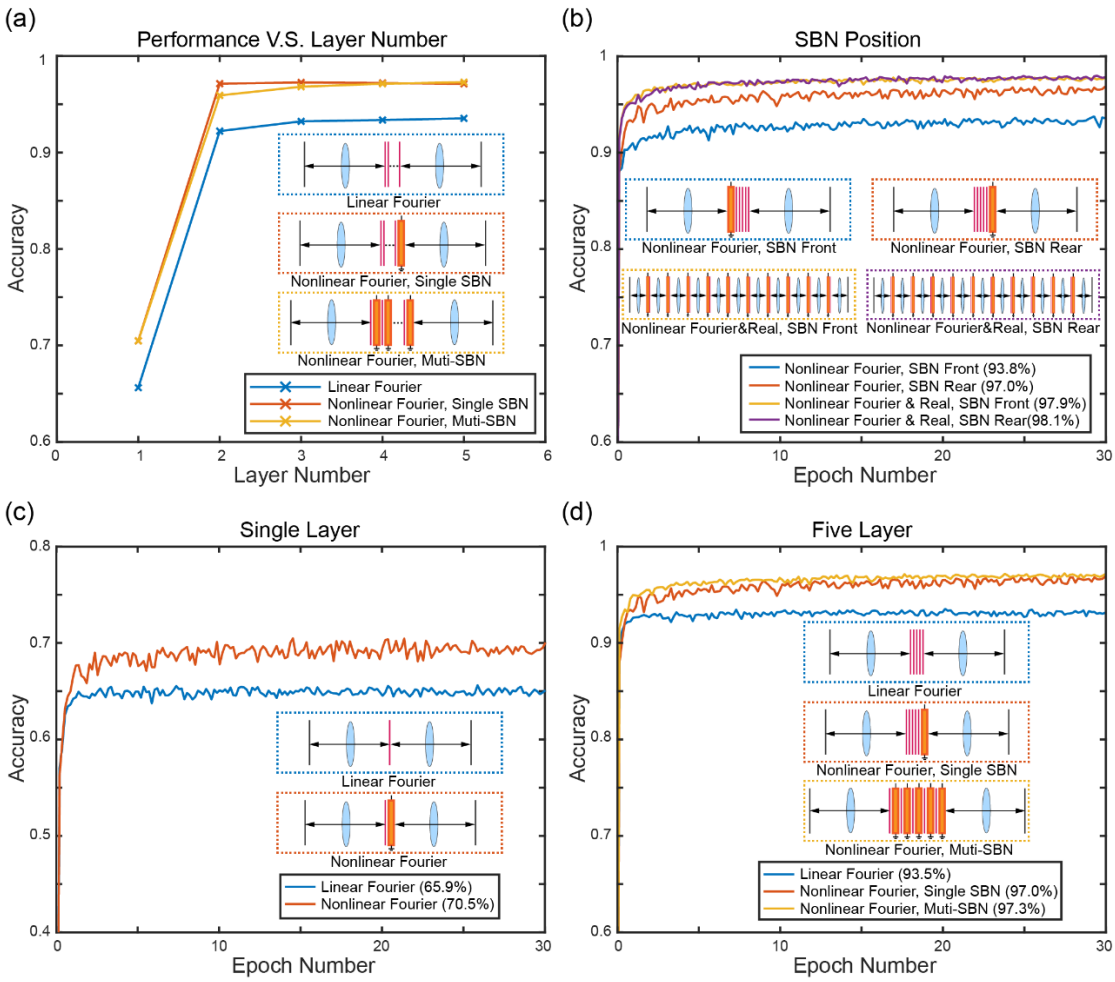



Supplementary Figure S5: Testing the Fourier-space $D^2NN$ on the benchmark dataset, i.e., Extended Complex Scene Saliency Dataset (ECSSD) [39]. We used the trained Fourier-space $D^2NN$ model of the main text for saliency detecting of 1,000 challenging natural images in ECSSD dataset. Since the network works at a single wavelength, all images were converted into grayscale as the input. Despite a lack of color information which is one of the major information the saliency detection based on, our physical network is still able to detect and segment out the saliency target (third column) for the input image (first column). We calculated and plotted the average precision-recall curve of the Fourier-space $D^2NN$ outputs over the whole dataset with respect to the ground truth saliency detection results (second column) as shown in the right half of the figure, and the maximum F-measure of the plot was calculated to be 0.423. The maximum F-measures of the RBD algorithm [29] and the state-of-the-art method [25] for saliency detection on ECSSD dataset with color information are 0.718 and 0.787, respectively. The numerical experimental results validated that the proposed approach can be used for saliency detecting of complicated natural images under different scenarios at the speed of light, and the performance of which can be further improved by using multiple wavelengths, adopting more advanced nonlinearity materials, and increasing the model complexity, etc.

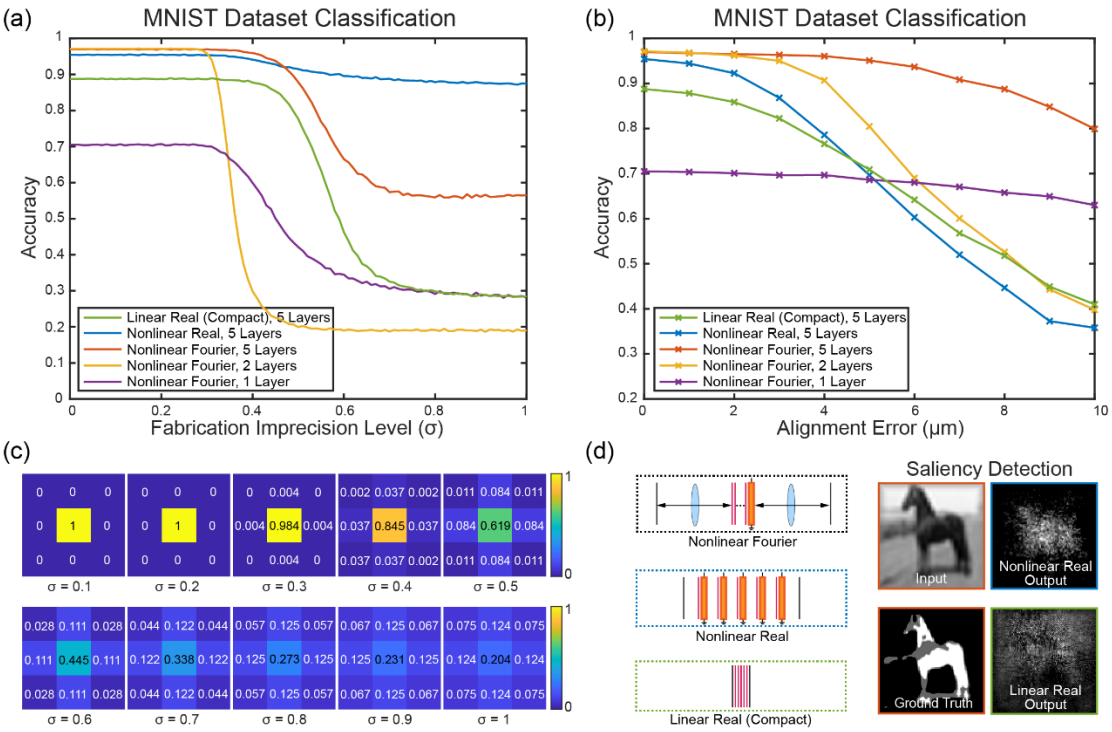



Supplementary Figure S6: The performance of F-D$^2$NN under the phase modulation range of $0\sim\pi$ and $0\sim2\pi$ on both tasks of saliency detection and classification. During the training, the phase modulation coefficients of the network were constrained within a certain range to facilitate the physical fabrication of modulation layers and provide enough degree-of-freedoms for training the target transfer function. For the saliency detection task in Fig. 2 of our main text, we found that constraining the phase modulation range to $0\sim\pi$ achieves the comparable performance with $0\sim2\pi$, as shown in the representative results in (a) and from the precision-recall curves in (b). The trained F-D$^2$NN modulation layers corresponding to the phase modulation range of $0\sim\pi$ and $0\sim2\pi$ are shown in (c), with which the maximum F-measure on the testing dataset are 0.616 and 0.613, respectively. The results of saliency detection task in this manuscript were trained with a phase modulation range of $0\sim2\pi$. For the task of MNIST dataset classification in Fig. 4 of our main text, we evaluated both linear and nonlinear Fourier-space D$^2$NN under the phase modulation range of $0\sim\pi$ and $0\sim2\pi$, and found that setting phase modulation range to $0\sim\pi$ instead of $0\sim2\pi$ improves the classification accuracy by 0.4%

and 1.5% for the linear and nonlinear Fourier-space $D^2NN$, respectively. The results of the classification task in this manuscript were trained with a phase modulation range of $0\sim\pi$. (Scale bar: 200 μm for (a) and 400 μm for (c). Cell scales are optically magnified.)

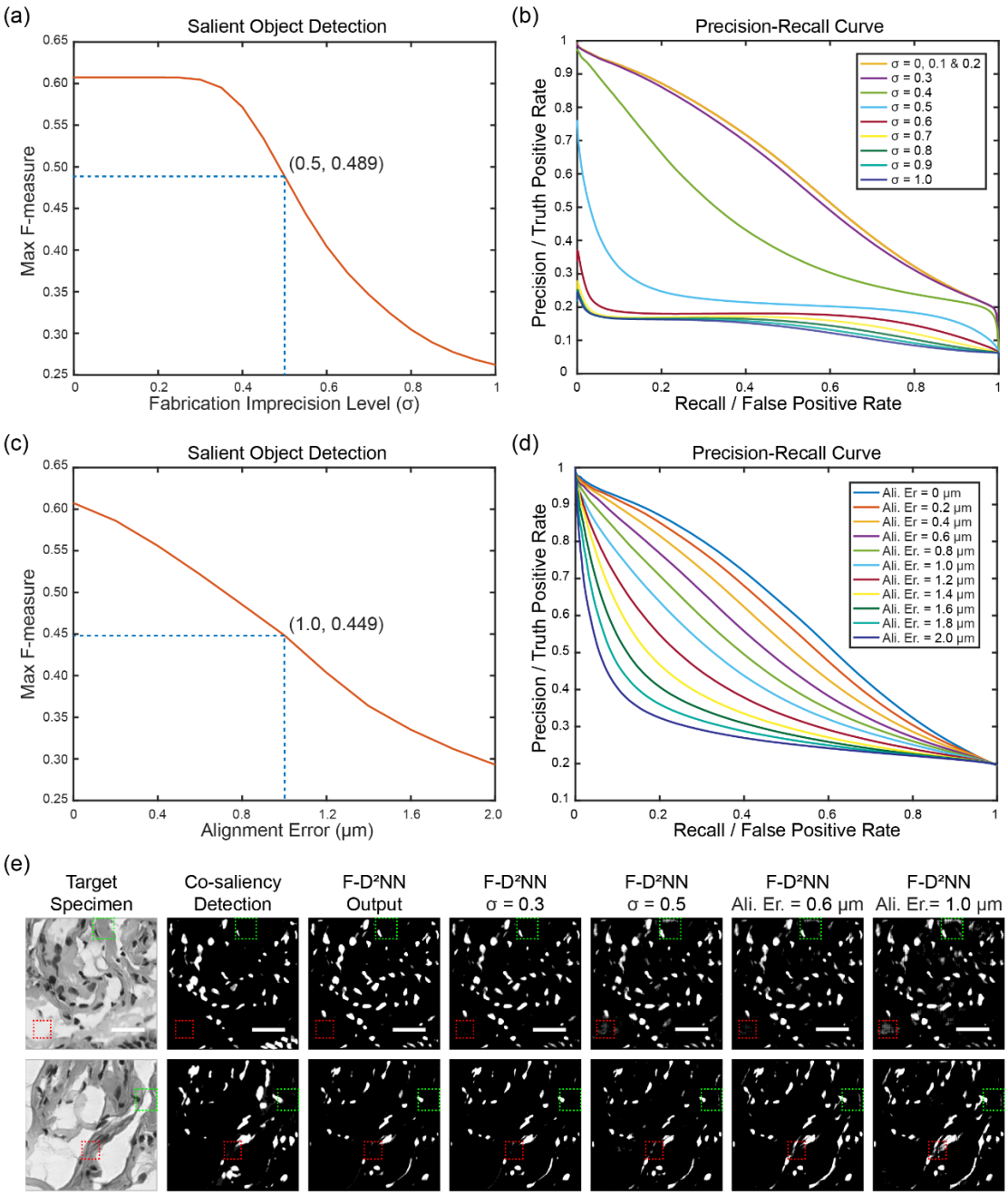



Supplementary Figure S7: The performance of the F-D$^2$NN with respect to the layer number and the position of the nonlinearity layer (photorefractive crystal, SBN:60) on the MNIST handwriting digits classification. (a) The increasing of layer number improves the accuracy of both nonlinear and linear F-D$^2$NN on classifying the MNIST dataset. Although nonlinear F-D$^2$NN significantly outperforms the linear F-D$^2$NN, using multiple nonlinear SBN layers can barely improve the classification accuracy but decrease the feasibility of physical implementation in only Fourier configuration. (b) Setting the nonlinear SBN layer behind modulation layers to perform the activation function improves the classification accuracy of both configurations, especially on the only Fourier configuration. (c, d) The convergence plots of single layer and five-layer F-D$^2$NNs. Compared with a single layer F-D$^2$NN, multilayer architecture provides higher degree-of-freedom to train the desired transfer function between the input and output planes, which achieves higher classification accuracy.




Supplementary Figure S8: The sensitivity of D$^2$NN with respect to the fabrication imprecision (a) and alignment error (b) under different configurations on the task of classifying MNIST dataset. We compare the sensitivity of Fourier space and real space D$^2$NN under different numbers of diffractive modulation layers (single-layer, two-layers, and five-layer). Three different types of configurations, including nonlinear Fourier-space, nonlinear real-space, and linear real-space configuration, are illustrated in the left of (d). The successive layer distance of the nonlinear Fourier space and linear real space D$^2$NN is set to be 100 um under the visible wavelength of 532 nm, with which the multiple diffractive layers of the network is considered to be fabricated with direct femtosecond laser writing all at once to avoid the layer to layer alignment. Thus, we consider the fabrication imprecision as the crosstalk between adjacent pixels during the fabrication by convolving a 3×3 Gaussian kernel with different standard deviations (σ), as shown in (c), to the trained diffractive modulation patterns during the testing. The alignment error is considered as the global shifting of the multiple diffractive layers. The results in (a) and (b) demonstrate that: (1) for nonlinear Fourier-space configuration, five-layer configuration has a significantly higher accuracy than single-layer configuration, and is also much more robust compared with two-layer configuration; (2) five-layer nonlinear Fourier-space D$^2$NN achieves higher accuracy and robustness than five-layer linear real-space configuration; (3) for five-layer nonlinear configurations, Fourier-space D$^2$NN achieves higher

accuracy and alignment robustness than real-space $D^2NN$, but real-space $D^2NN$ is more robust to fabrication imprecision. Besides, different from Fourier-space $D^2NN$, real-space $D^2NN$ cannot perform saliency detection, as shown in (d).




Supplementary Figure S9: The performance of the five-layer nonlinear F-D$^2$NN sensitivity with respect to the fabrication imprecision (a, b) and alignment error (c, d) on the task of saliency detection. Again, we consider the fabrication imprecision as the crosstalk between adjacent pixels during the direct laser writing by convolving a 3×3 Gaussian kernel with different standard deviations (σ) to the trained diffractive modulation patterns during the testing. The alignment error is considered as the global shifting of the diffractive layers. The robustness of five-layer nonlinear F-D$^2$NN is evaluated by training it on the same pathology slide datasets (Cell Type 1) we used in the main text. We plotted the average precision-recall curves of the

saliency detection results on testing datasets with different amount of the fabrication and alignment error, as shown in (b) and (d), respectively; the corresponding maximum F-measures are calculated and plotted in (a) and (c), respectively. The plots in (a) and (c) demonstrate that although the performance of saliency detection decreases with the increasing of fabrication and alignment errors, even under the standard deviation of 0.5 as the fabrication imprecision or pixel shifting of 1.0 μm as the alignment error, the five-layer nonlinear F-$D^2$NN is still able to perform decent saliency detection with maximum F-measures of 0.489 and 0.449, respectively. The representative saliency detection results on target specimen (Cell Type 2) under the standard deviation of 0.3 and 0.5 as well as pixel shifting of 0.6 μm and 1.0 μm are demonstrated in (e). (Scale bar: 200 μm. Cell scales are optically magnified.)