Misalignment Resilient Diffractive Optical Networks

Deniz Mengu1,2,3, Yifan Zhao1,3, Nezih T. Yardimci1,3, Yair Rivenson1,2,3, Mona Jarrahi1,3, Aydogan Ozcan1,2,3,4,*

1 Electrical and Computer Engineering Department, University of California, Los Angeles, CA, 90095, USA
2 Bioengineering Department, University of California, Los Angeles, CA, 90095, USA
3 California NanoSystems Institute, University of California, Los Angeles, CA, 90095, USA
4 Department of Surgery, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095, USA.
* Corresponding author: ozcan@ucla.edu

Supplementary Information (SI)

1 Experimental Setup Details

The schematic diagram of the experimental setup is given in Fig. 5E of the main text. The THz wave incident on the object was generated through a horn antenna compatible with the source WR2.2 modular amplifier/multiplier chain (AMC) from Virginia Diode Inc. (VDI). Electrically modulated with 1 kHz square wave, the AMC received an RF input signal that is a 16 dBm sinusoidal waveform at 11.111 GHz ($f_{RF1}$). This RF signal is multiplied 36 times to generate the continuous-wave (CW) radiation at 0.4 THz, corresponding to ~0.75 mm in wavelength. The exit aperture of the horn antenna was placed ~60 cm away from the object plane of the 3D-printed diffractive optical network. At the output plane of the diffractive optical network, we 3D-printed an output aperture that has 10 openings, each with a size of 4.8 mm × 4.8 mm, defining the class detectors at their relative locations. The diffracted THz light at the output plane was collected with a single-pixel Mixer/AMC from Virginia Diode Inc. (VDI). A 10 dBm sinusoidal signal at 11.083 GHz was sent to the detector as local oscillator for mixing, and the down-converted signal was at 1GHz. The 10 openings representing the class detectors was scanned by placing the single-pixel detector on an XY stage that was built by combining two linear motorized stages (Thorlabs NRT100). The scanning step size was set to be 1 mm within each aperture opening. The down-converted signal of single-pixel detector at each scan location was sent to low-noise amplifiers (Mini-Circuits ZRL-1150-LN+) to amplify the signal by 80 dBm and a 1 GHz (+/-10 MHz) bandpass filter (KL Electronics 3C40-1000/T10-O/O) to clean the noise coming from unwanted frequency bands. Following the amplification, the signal was passed through a tunable attenuator (HP 8495B) and a low-noise power detector (Mini-Circuits ZX47-60), then the output voltage was read by a lock-in amplifier (Stanford Research SR830). The modulation signal was used as the reference signal for the lock-in amplifier and accordingly, we conducted a calibration by tuning the attenuation and record the lock-in amplifier readings. The lock-in amplifier readings at each scan location were converted to linear scale according to the calibration. The class scores shown in Fig. 6 of the main text as well as the Supplementary Figs. S2 and S3, were computed as the sum of the calibrated and converted lock-in amplifier output at every scan step within the corresponding class detector opening.

The diffractive optical networks were fabricated using a 3D printer (Objet30 Pro, Stratasys Ltd.). Each 3D-printed diffractive optical network consisted of an input object, 5 diffractive layers and an output aperture array corresponding to the desired locations of the class detectors (see Fig. 1A of main text). While the active modulation area of our 3D printed diffractive layers was 8 cm × 8 cm (106.66a × 106.66a), they were printed as light modulating insets surrounded by a uniform slab of printed material with a thickness of 0.9 mm. The total size of each printed layer was selected large enough to accommodate the introduced shifts on the 3rd diffractive layer location (for misalignment testing), with a total size of 12.8 cm × 12.8 cm.
The output aperture array and the 3D-printed MNIST digits were coated with aluminum except the openings and object features. Each aperture at the output plane is a square covering an area of 4.8 mm × 4.8 mm, matching the assumed size during the training. The size of the printed MNIST digits was 4 cm × 4cm sampled at a rate of 0.4 mm in both x and y directions, matching the training forward model. A 3D-printed holder was used to align the 3D printed input object, 5 diffractive layers and the output aperture. Around the location of the 3rd layer, the holder had additional spatial features that allowed us to move this diffractive layer to 13 different locations including the ideal one (see Fig. 5 of the main text).

2 Training of diffractive optical networks

2.1 Forward optical model

2.1.1 Error-free D^3NN

In a diffractive optical network, each unit diffractive feature of a layer represents a complex-valued transmittance learned during the training process that optimizes the thickness, h, of the features based on the complex-valued refractive index of the 3D-fabrication material, τ = n + jκ. The characterization of the printing material in a THz-TDS setup revealed the values of n and κ as 1.7227 and 0.031, respectively, for a monochromatic THz light at 400 GHz. Our formulation represents the complex-valued transmittance function of a diffractive feature on layer, l, at coordinates (x_q, y_q, z_l) as:

\[ t(x_q, y_q, z_l) = \exp\left( -\frac{2\pi \kappa h(x_q, y_q, z_l)}{\lambda} \right) \exp\left( j\frac{2\pi(n-n_{air}) h(x_q, y_q, z_l)}{\lambda} \right) \]  

(S1),

where \( h(x_q, y_q, z_l) \), \( n_{air} \) and \( z_l \) denote the thickness of a given feature, refractive index of air and the axial location of the layer, l, respectively. From the Rayleigh-Sommerfeld theory of diffraction, we can interpret every diffractive unit on layer, l, at (x_q, y_q, z_l), as the source of a secondary wave, \( w_l^q(x, y, z) \),

\[ w_l^q(x, y, z) = \frac{1}{r^2} \exp\left( j\frac{2\pi r}{\lambda} \right) \]  

(S2),

where \( r = (\sqrt{(x-x_q)^2 + (y-y_q)^2} + (z-z_l)^2)^{0.5} \). Therefore, the complex field coming out of the qth feature of (l+1)th layer, \( u_{l+1}^q(x, y, z) \) can be written as:

\[ u_{l+1}^q(x, y, z) = t(x_q, y_q, z_{l+1}) w_{l+1}^q(x, y, z) \sum_{k \in l} w_k^q(x_q, y_q, z_{l+1}) \]  

(S3).

We sampled our diffractive fields and surfaces at a sampling interval of 0.4 mm that is equal to 0.53λ. The smallest diffractive feature size was also equal to 0.4 mm. The learnable thickness of each feature, h, was defined over an auxiliary variable, \( h_a \):

\[ h = q\left( \frac{\sin(h_a)+1}{2} \right) + h_b \]  

(S4),

where \( h_m \) and \( h_b \) denote the maximum modulation thickness and base thickness, respectively. Taking \( h_b \) as 0.5 mm and \( h_m \) as 1 mm, we limited the printed thickness values between 0.5 mm and 1.5 mm. The minimum thickness \( h_b \) was used to mainly ensure the mechanical stability of the 3D printed layers against cracks and bending. The operator q(.) in Eq. (S4) represents the quantization operator. We quantized the thickness values to 16 discrete levels (0.0625 mm per step). For the initialization of the diffractive layers at the beginning of the training, the thickness of each feature was taken as a uniformly distributed random variable between 0.9 mm and 1.1 mm, including the base thickness.

2.1.2 Vaccinated D^3NN

The training of the vaccinated diffractive optical networks follows the same optical forward model outlined in the previous section, except that it additionally introduces statistical variations following the models of the error sources in
a diffractive network. The components of the 3D displacement vector of the l\textsuperscript{th} diffractive layer, $\mathbf{D}^l = (D^l_x, D^l_y, D^l_z)$, were defined as uniformly distributed random variables defined by Eq. (1) of the main text. The vaccination strategy uses different sets of displacement vectors at every iteration (batch) to introduce undesired misalignments of the diffractive layers during the training. With $\mathbf{D}^{(l,i)} = (D^{(l,i)}_x, D^{(l,i)}_y, D^{(l,i)}_z)$ denoting the random displacement that the l\textsuperscript{th} layer experiences at i\textsuperscript{th} iteration, Eq. (S3) was adjusted according to the longitudinal shift of the successive layers, $D^l_x(l,i)$ and $D^l_z(l+1,i)$, i.e., the light propagation distances between the diffractive layers were varied at every iteration. To implement the continuous lateral displacement of diffractive layers, we used:

$$t^{(l,i)}(x+D^{(l,i)}_x, y+D^{(l,i)}_y) = \int \int T^{(l,i)}(u, v) \exp(j2\pi (u(x+D^{(l,i)}_x) + v(y+D^{(l,i)}_y))) \, du \, dv$$  \hspace{1cm} (S5),

where $t^{(l,i)}(x,y)$ denotes the 2-dimensional complex modulation function of layer l, at i\textsuperslash{}th iteration, and $T^{(l,i)}(u,v)$ represents its spatial Fourier transform defined over the 2D spatial frequency space (u,v).

### 2.2 Training of all-optical and hybrid classification systems

#### 2.2.1 Loss function and class scores

In our forward training model, without loss of generality, we modeled our detectors as ratiometric sensors that capture the ratio of the optical power incident over their active area, $P_d$, and the optical power incident over the object at the input plane, $P_{obj}$. Based on this, the optical signal vector collected by output detectors, $\mathbf{I}_d$, was formulated as:

$$\mathbf{I}_d = \frac{P_d}{P_{obj}}$$  \hspace{1cm} (S6).

For all three diffractive object classification systems depicted in Fig. 1 of main text, the cost function was defined as the widely-known softmax-cross-entropy (SCE):

$$\mathcal{L} = - \sum_{c=1}^{C} g_c \log\left(\frac{\exp(s_c)}{\sum_{c=1}^{C} \exp(s_c)}\right)$$  \hspace{1cm} (S7),

where $g_c$, $s_c$ and C denote the binary entry in the label vector, the computed class score for the data class, c, and the number of data classes in a given dataset (e.g., C=10), respectively.

For the standard diffractive optical network architecture shown in Fig. 1A, the number of class detectors, $N_d$, is equal to the number of data classes, C. In this scheme, the class score vector, $\mathbf{s}$, was computed by:

$$s = T \cdot \frac{\mathbf{I}_d}{\max(\mathbf{I}_d) + \varepsilon}$$  \hspace{1cm} (S8),

where $T$ and $\varepsilon$ are constants, i.e., non-trainable hyperparameters used during the training phase. The multiplicative factor $T$ was empirically set to be equal to 10 to generate artificial signal contrast at the input of softmax function for more efficient convergence of training. The constant $\varepsilon$, on the other hand, was used to regularize the power efficiency of the standard diffractive object recognition systems. In particular, the standard diffractive neural network models presented in Figs. 2A, 2D and 3 of the main text, as well as in the Supplementary Figs. S4A, S4D and S5, were trained by taking $\varepsilon = 10^{-4}$ which results in low power efficiency, $\eta$, and low signal contrast, $\psi$. The 3D-printed diffractive optical networks, on the other hand, were trained by setting $\varepsilon = 10^{-3}$ to circumvent the effects of the limited signal-to-noise ratio in our experimental system. Trained with a higher $\varepsilon$ value, these diffractive networks offer slightly compromised blind testing accuracies while providing significantly improved power efficiency, $\eta$, and signal contrast, $\psi$, which are defined as:

$$\eta = I_{gt}$$  \hspace{1cm} (S9a),

$$\psi = I_{gt} - I_{sc}$$  \hspace{1cm} (S9b),
where $I_{gt}$ and $I_{sc}$ denote the optical signals measured by the class detector representing the ground truth label of the input object and its strongest competitor, i.e. the second maximum for a correctly classified input object, respectively. A comparison between the inference performances of low- and high-contrast variants of vaccinated and non-vaccinated standard diffractive optical networks under various levels of misalignments is presented in Fig. S1. As depicted in Fig. S1A, the high contrast, high efficiency standard diffractive networks are more robust against the undesired system variations/misalignments compared to their low-efficiency counterparts when both networks were trained under error-free conditions. Figure S1B, on the other hand, compares the standard diffractive network architectures that were tested within the same misalignment range used in their training. In this case, the low-contrast, power inefficient diffractive networks show their higher inference capacity advantage and adapt to the misalignments more efficiently than the diffractive classification systems trained to favor higher power efficiency.

In a differential diffractive optical network system, the number of detectors is doubled, i.e. $N_d=2C$, where each pair represents the negative, $I_{d-}$, and positive signal vector, $I_{d+}$, contributing to the normalized differential signal, $I_{(d,n)}$ (see Fig. 1B of the main text) defined as:

$$I_{(d,n)} = \frac{I_{d+} - I_{d-}}{I_{d+} + I_{d-}}. \quad (S10)$$

In parallel, the class scores of a differential diffractive object classification system, $s$, are calculated by replacing the optical signal vector, $I_d$, in Eq. (S8) with the normalized differential signals, $I_{(d,n)}$, depicted in Eq. (S10).

It is important to note that the Eqs. (S6), (S7) and (S8) concern only the training stage of diffractive optical networks and the associated all-optical object classification systems. Once the training is completed, these equations are not used in the numerical and experimental blind testing, meaning that the class decision is made solely based on $\max(P_d)$ and $\max(P_{(d,n)})$ in standard and differential diffractive network systems, respectively.

In the hybrid neural network models, we jointly-trained 5-layer diffractive optical networks with an electronic network that has a single-layer fully-connected network with only 110 (100 multiplicative weights + 10 bias) trainable parameters. During the joint-evolution of these two neural networks, we normalized the optical signal collected by the detectors, $I_d$, as depicted in Eq. (S8) with $T = 1$. These normalized detector signals were then fed into the subsequent fully-connected layer in the electronic domain to compute the class scores, $s$, which was used in Eq. (S7) for computing the classification loss before the error-backpropagation through both the electronic and diffractive optical networks.

2.2.2 Other training related details

All network models used in this work were trained using Python (v3.6.5) and TensorFlow (v1.15.0, Google Inc.). We selected Adam optimizer during the training of all the models, and its parameters were taken as the default values in TensorFlow and kept identical in each model. The learning rates of the diffractive optical networks and the electronic neural network were set to be 0.001 and 0.0002, respectively. The data of handwritten digits and fashion-products were both divided into three parts: training, validation and testing, containing 55K, 5K and 10K images, respectively. All object recognition systems were trained for 50 epochs with a batch size of 50 and the best model was selected based on the highest classification performance on the validation dataset. In the training of MNIST digits, the image information was encoded in the amplitude channel at the object plane, while the Fashion-MNIST objects was assumed to be phase-only targets with their gray levels mapped to phase values between 0 and $\pi$. 
Supplementary Figures

Fig. S1: The comparison between the low-contrast and high-contrast standard diffractive optical networks. A The inference accuracy values of two error-free standard optical network designs are compared. The low-contrast standard diffractive optical network (red) achieves slightly higher inference accuracy when the alignment is perfect. The high-contrast diffractive optical network, on the other hand, is slightly more robust against misalignments. B Trained with the ν-D^NN framework, low-contrast models use their higher inference capacity to adapt to misalignments, consistently achieving higher classification accuracies when they are tested under misalignment.
Fig. S2: Experimental image classification results as a function of misalignments. A The experimentally measured class scores for handwritten digit ‘0’ selected from Set 2. B Same as A, except the input object is now a handwritten digit ‘7’ selected from Set 2. The red dot within the coordinate system shown on the left-hand side represents the physical misalignment for each case (see Fig. 5B of main text). Red (green) rectangles mean incorrect (correct) inference results.
Fig. S3: Experimental image classification results as a function of misalignments. A The experimentally measured class scores for handwritten digit '2' selected from Set 1. B Same as A, except the input object is now a handwritten digit '3' selected from Set 2. The red dot within the coordinate system shown on the left-hand side represents the physical misalignment for each case (see Fig. 5B of main text). Red (green) rectangles mean incorrect (correct) inference results.
Fig. S4: The blind inference accuracies achieved by standard, differential and hybrid diffractive network systems for the classification of phase-encoded Fashion-MNIST images. Same as the Figure 2 of main text, except, the image dataset is Fashion-MNIST. Unlike amplitude encoded MNIST images at the input plane, the fashion products were assumed to represent phase-only targets at the object/input plane with their phase values restricted between 0 and π.
Fig. S5: Direct comparison of blind inference accuracies achieved by standard, differential and hybrid diffractive network systems for the classification of phase-encoded fashion products. Same as the Figure 3 of main text, except, the image dataset is Fashion-MNIST. Unlike amplitude encoded MNIST images at the input plane, the fashion products were assumed to represent phase-only targets at the object/input plane with their phase values restricted between 0 and π.