Time-Lapse Image Classification Using a Diffractive Neural Network

*Md Sadman Sakib Rahman and Aydogan Ozcan**

Diffractive deep neural networks (D^{2}NNs), comprised of spatially engineered passive surfaces, collectively process optical input information at the speed of light propagation through a thin diffractive volume, without any external computing power. Diffractive networks were demonstrated to achieve all-optical object classification and perform universal linear transformations. Herein, a "time-lapse" image classification scheme using a diffractive network is demonstrated for the first time, significantly advancing its classification accuracy and generalization performance on complex input objects by using the lateral movements of the input objects and/or the diffractive network, relative to each other. In a different context, such relative movements of the objects and/or the camera are routinely being used for image super-resolution applications; inspired by their success, a time-lapse diffractive network is designed to benefit from the complementary information content created by controlled or random lateral shifts. The design space and performance limits of time-lapse diffractive networks are numerically explored, revealing a blind testing accuracy of 62.03% on the optical classification of objects from the CIFAR-10 dataset. This constitutes the highest inference accuracy achieved so far using a single diffractive network on the CIFAR-10 dataset. Time-lapse diffractive networks will be broadly useful for the spatiotemporal analysis of input signals using all-optical processors.

1. Introduction

Machine learning and artificial intelligence research has experienced rapid growth in the last two decades.[1] One of the core engines that has driven this growth is deep learning,[2] permitting efficient and rapid training of deep artificial neural network models. The ability to train deep neural networks has revolutionized artificial intelligence, and electronics has been the undisputed platform of choice for implementing artificial neural networks. Specialized processing hardware such as graphics processing unit (GPU) is widely used today for deep learning. However, these electronic processors are power-hungry and bulky, making researchers wary of the environmental impact of machine learning.[3,4] Therefore, there is strong interest in low-power and fast computing platforms for machine learning applications. Optical computing has been identified as a promising potential alternative for such purposes because of the large bandwidth, high speed, and massive parallelism of optics.[5]

Diffractive deep neural networks (D^{2}NNs), also known as diffractive optical networks or diffractive networks, form a passive all-optical computing platform that exploits the diffraction of light waves to perform computation.[6] These diffractive networks are composed of several spatially engineered surfaces, separated by free-space. The diffractive features/elements of a layer, also termed "diffractive neurons", locally modulate the amplitude and/or the phase of the light incident upon the layer. Successive modulation by and diffraction through the layers give rise to an all-optical transformation between the input and the output fields-of-view at the speed of light propagation without any external power. The amplitude and/or the phase values of the diffractive neurons corresponding to a desired optical transformation or computational task are trained/learned through a digital computer using deep learning. Once the training is complete, the layers can be fabricated and assembled to form a "physical" network that performs the desired computation in a passive manner and at the speed of light propagation. Diffractive networks can achieve universal linear transformations,[7–9] and various applications using diffractive processors have been demonstrated such as object classification, pulse processing, imaging through random diffusers, hologram reconstruction, quantitative phase imaging, class-specific imaging, super-resolution image display, all-optical logic

M. S. S. Rahman, A. Ozcan
Electrical and Computer Engineering Department
University of California Los Angeles
Los Angeles, CA 90095, USA
E-mail: ozcan@ucla.edu

M. S. S. Rahman, A. Ozcan
Bioengineering Department
University of California Los Angeles
Los Angeles, CA 90095, USA

M. S. S. Rahman, A. Ozcan
California NanoSystems Institute (CNSI)
University of California Los Angeles
Los Angeles, CA 90095, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202200387.

DOI: 10.1002/aisy.202200387

operations, beam shaping, and orbital angular momentum mode processing, among others.[10–30]

While diffractive networks have shown competitive performance on the classification of relatively simpler objects, for example, handwritten digits and fashion products,[11] for more complex natural objects such as those from the CIFAR-10 dataset,[31] their performance gap compared to the classification accuracy of electronic neural networks is still large.[11,32] Ensemble learning through multiple D^2NNs has been demonstrated to improve the inference and generalization of diffractive networks at the cost of reducing the compactness and simplicity of the optical hardware.[32]

In this work, we demonstrate, for the first time, a "time-lapse" image classification scheme with a stand-alone diffractive optical network that significantly enhances the inference and generalization performance of diffractive computing. In this scheme, the objects and/or the diffractive network laterally move relative to each other, either randomly or in a controlled manner, during the detector integration time, enriching the information provided to the diffractive network. In a different context and application, lateral shifts of the object of interest relative to the imager have been routinely used for pixel super-resolution imaging, enhancing the resolution of the reconstructed images.[33–36] Inspired by the success of these pixel super-resolution approaches, here we use the controlled or random relative displacements between the input objects and the diffractive network for time-lapse image classification and report a numerical blind testing accuracy of 62.03% for the classification of grayscale CIFAR-10 images, which constitutes the highest classification accuracy for this dataset achieved so far using a single diffractive optical network. In addition to significantly advancing the inference and generalization performance of D^2NNs, these time-lapse diffractive networks can also find broader use in the all-optical processing of spatiotemporal information of a scene or object.

2. Results

The concept of time-lapse image classification with a diffractive network is illustrated in **Figure 1**. A diffractive network comprising five phase-only diffractive layers, axially separated by 40λ, is placed between the object plane and the detector plane. The detector plane includes 20 detectors:[11] 2 detectors for each class c of the CIFAR-10 dataset, i.e., a "positive" detector $D_{c,+}$ and a "negative" detector $D_{c,-}$. The integration time of the output detectors is assumed to be $N\delta_t$, where N is the number of lateral object shifts and each of the N individual shifts has an equal

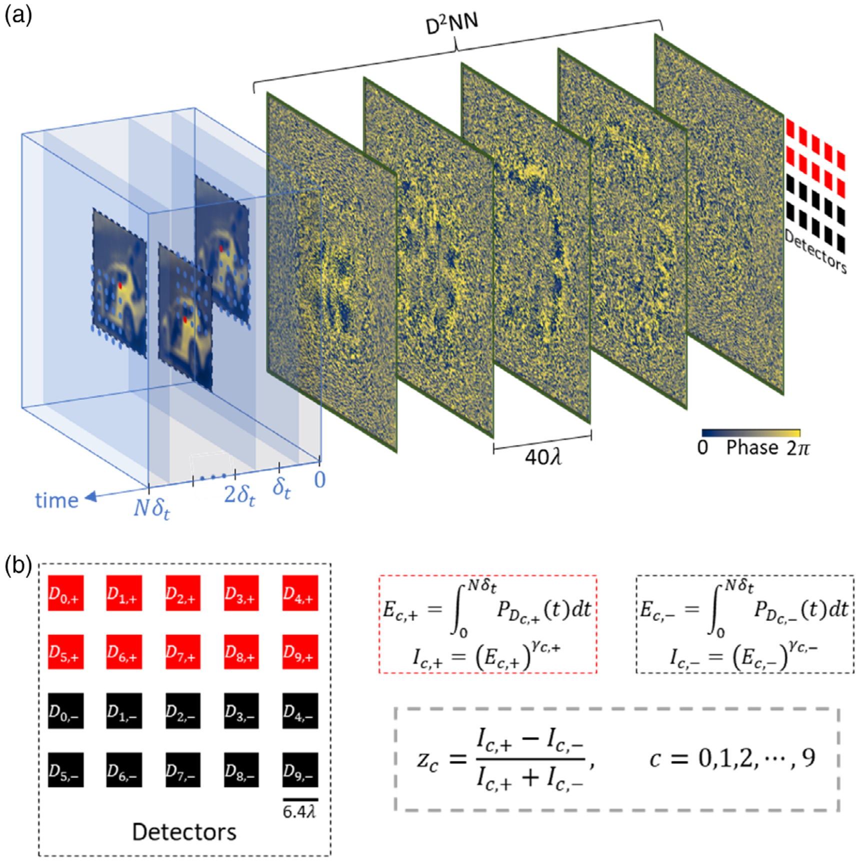
(a)
D²NN
Detectors
40λ
0 Phase 2π
time
Nδ_t
2δ_t
δ_t
0
(b)
D_{0,+} D_{1,+} D_{2,+} D_{3,+} D_{4,+}
D_{5,+} D_{6,+} D_{7,+} D_{8,+} D_{9,+}
D_{0,−} D_{1,−} D_{2,−} D_{3,−} D_{4,−}
D_{5,−} D_{6,−} D_{7,−} D_{8,−} D_{9,−}
6.4λ
Detectors
$E_{c,+} = \int_0^{N\delta_t} P_{D_{c,+}}(t)dt$
$I_{c,+} = (E_{c,+})^{\gamma_{c,+}}$
$E_{c,-} = \int_0^{N\delta_t} P_{D_{c,-}}(t)dt$
$I_{c,-} = (E_{c,-})^{\gamma_{c,-}}$
$z_c = \frac{I_{c,+} - I_{c,-}}{I_{c,+} + I_{c,-}}, \quad c = 0,1,2,\cdots,9$

Figure 1. Time-lapse image classification using a D^2NN. a) A diffractive network with five phase-only diffractive layers followed by 20 detectors at the detector plane for differential image classification. The integration time of each detector is $N\delta_t$. During each one of the N intervals of δ_t duration, the center of the object is laterally displaced to a new point (red circle); these lateral displacements can be entirely random or follow a predefined grid (blue circles). b) Labeling of the detectors where $D_{c,+}$ ($D_{c,-}$) denote the positive (negative) detectors assigned to class c. The differential class scores z_c are used for the final classification decision based on the maximum score.

integration time of δ_t. Without changing our conclusions, in alternative implementations, the diffractive network can also laterally move relative to the static object, or both the object and the diffractive network can laterally move at the same time. Each detector $D_{c,\pm}$ is assigned an exponent $\gamma_{c,\pm}$ which operates on the integrated detector power to yield the detector signal $I_{c,\pm}$ (see Figure 1 and the Experimental Section). We will report diffractive classification results under two different conditions: 1) the exponents are assumed to be trainable, and 2) nontrainable, fixed as $\gamma_{c,\pm} = 1$. The normalized differential class scores $z_c = \frac{I_{c,+} - I_{c,-}}{I_{c,+} + I_{c,-}}$ are calculated from these detector signals, and the prediction/inference is made in favor of the class receiving the highest differential optical score (see Figure 1).

For all the D^{2}NNs reported in this work, each trainable diffractive layer consists of 200 × 200 diffractive elements (diffractive neurons) of size $0.53\lambda \times 0.53\lambda$. The objects are assumed to be phase-only and the diffractive networks are trained using the grayscale CIFAR-10 dataset (refer to the Experimental Section for details). The hyperparameters that define the grid of lateral displacements of the objects during the time-lapse image classification are $s_{\max}$ and m, where $s_{\max}$ is the maximum (relative) lateral displacement along x/y and m^2 refers to the total number of points on the grid (see **Figure 2**). The size of the input aperture is another hyperparameter that affects the classification performance of the time-lapse diffractive networks. The impact of these hyperparameters, $s_{\max}$, m, and the input aperture size, on the performance of time-lapse diffractive classifiers is shown in Figure 2. The classification performance is quantified by the blind testing accuracy of the networks on 10 000 previously unseen images belonging to the test set of the CIFAR-10 dataset. To obtain each data point in Figure 2, we trained three different diffractive networks with the same hyperparameters and calculated the mean and standard deviation of blind testing accuracies of these three trained networks. We see from Figure 2a that as $s_{\max}$ is increased from 3.20λ to 6.40λ (while keeping $m = 5$ and the aperture size $= 44.8\lambda \times 44.8\lambda$ constant), the mean blind testing accuracy increases until $s_{\max} = 5.33\lambda$, where it reaches its highest value of 61.35%. Beyond $s_{\max} = 5.33\lambda$, the mean classification accuracy starts to decrease. In Figure 2b, we set $s_{\max} = 5.33\lambda$, aperture size $= 44.8\lambda \times 44.8\lambda$, and vary m. As m is varied between 3 and 6, the mean accuracy increases rapidly from 58.56% to 61.35% until $m = 5$, beyond which the mean accuracy reaches a plateau. For Figure 2c, we selected $m = 5$ and $s_{\max} = 5.33\lambda$ (as optimized from Figure 2a,b) and the width of the input aperture was varied between 32.0λ and 53.3λ. The highest mean accuracy (Figure 2c) is observed for an input aperture size of $38.4\lambda \times 38.4\lambda$, which is smaller than the object support $44.8\lambda \times 44.8\lambda$. We compared this observation with its counterpart for time-static diffractive image classification

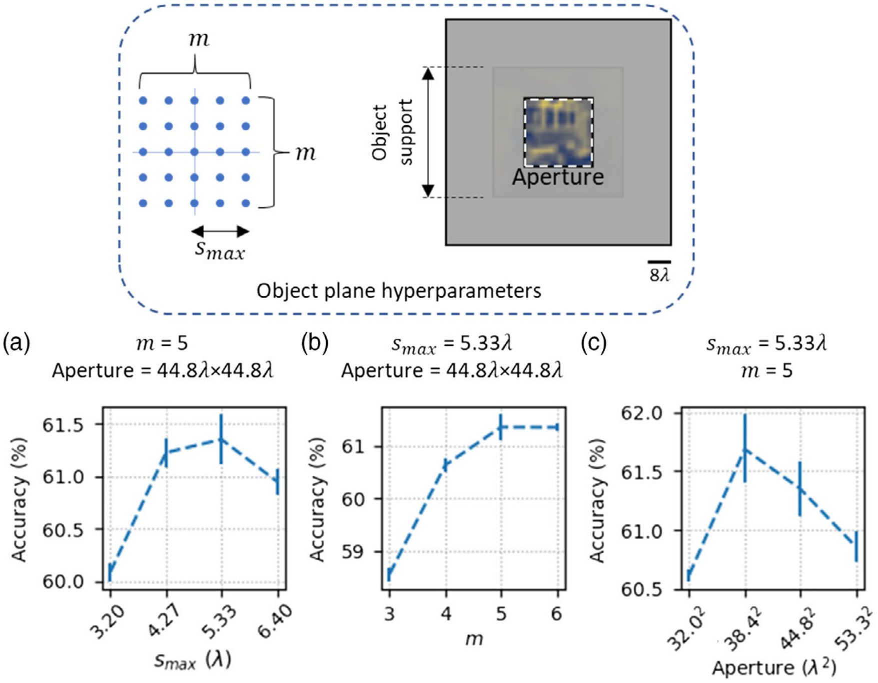

Figure 2. Dependence of the time-lapse image classification D^{2}NN performance on the hyperparameters of the input plane. Top left: Grid of points representing the lateral displacements of the input objects during the training, defined by the two hyperparameters, $s_{\max}$ and m. $s_{\max}$ represents the maximum displacement along either the vertical or the horizontal direction, whereas m^2 is the total number of grid points. Top right: the dashed white square represents the input aperture immediately following the object, the area of which is another hyperparameter. a) Dependence of the blind testing accuracy on $s_{\max}$ with m and the input aperture kept constant. b) Effect of m on the blind testing accuracy of the trained time-lapse diffractive classifiers as s_{max} and the input aperture are kept constant. c) Dependence of the blind testing accuracy on the input aperture size while $s_{\max}$ and m are kept constant. For (a)–(c), the data points and the error bars represent the mean and the standard deviation values, respectively, calculated from three designs, which are obtained by training three different time-lapse D^{2}NN classifiers for the same set of hyperparameter values. The curves are linearly interpolated between the data points.

(see Table S1, Supporting Information), where the aperture size corresponding to the highest mean blind testing accuracy is larger than the object support. This comparison indicates that a time-lapse diffractive network prefers a relatively smaller input field-of-view compared to its time-static counterparts.

Next, we juxtapose a time-lapse image classification diffractive network with a time-static diffractive network (see **Figure 3**). For this comparison, we chose the time-lapse diffractive network with the best individual blind testing accuracy (62.03%) among the networks constituting the results of Figure 2 and the time-static diffractive network with the best individual blind testing accuracy (53.14%) among the networks constituting the results of Table S1, Supporting Information. For the time-lapse image classification diffractive network, the hyperparameters

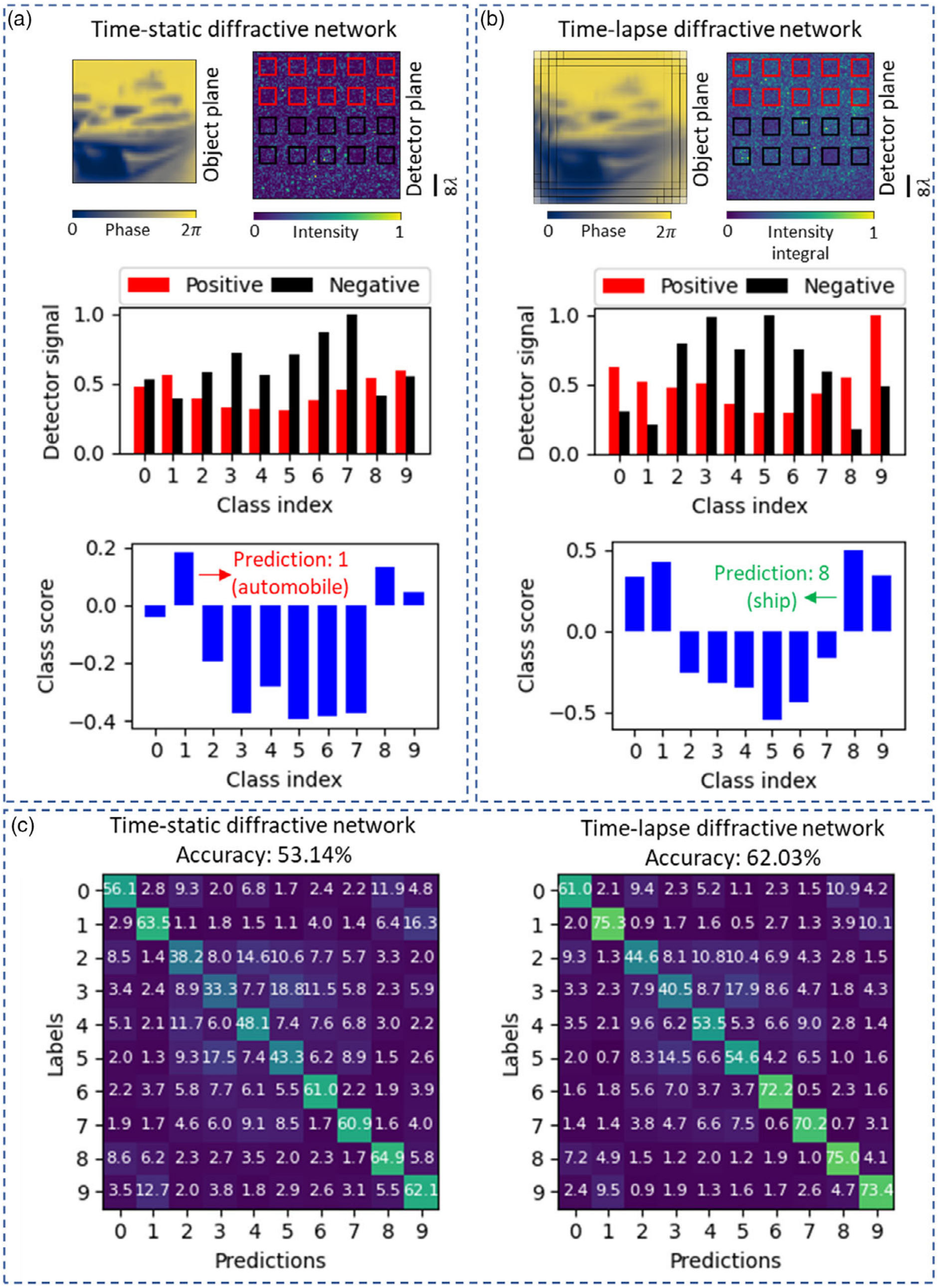

Figure 3. Comparison between a time-static and a time-lapse diffractive image classifier. a) Detector plane intensity, detector signals, and class scores from the time-static network for an object (a ship) from the data class 8 of the CIFAR-10 dataset. Based on the class scores, the object is misclassified to be an automobile. b) Integrated detector plane intensity, detector signals, and class scores from the time-lapse diffractive network for the same object, which is correctly classified as a ship. c) Confusion matrices for the two networks, evaluated by blind testing on 10 000 CIFAR-10 test images. The overall accuracies of the networks are 53.14% and 62.03%, respectively. The trained parameters of these two networks are shown in Figure S2a,b, Supporting Information.

corresponding to the highest individual accuracy were $m = 5$, $s_{\max} = 5.33\lambda$, and input aperture size $= 38.4\lambda \times 38.4\lambda$; while for the time-static network, the input aperture size corresponding to best individual accuracy was $51.2\lambda \times 51.2\lambda$. Another difference to be noted between the time-static and the time-lapse diffractive networks chosen for comparison in Figure 3 is that for the time-static one, the detector exponents were not trainable, i.e., $\gamma_{c,\pm} = 1$, whereas the detector exponents were trainable for the time-lapse network. The reason for this selection is that, unlike the time-lapse diffractive networks, time-static diffractive networks showed overfitting when the detector exponents are trainable, leading to inferior generalization (see Table S2, Supporting Information).

For an example object from the image class "ship" (true label: 8), we show in Figure 3a the detector plane intensity, detector signals, and the class scores for the time-static network; similarly, in Figure 3b we show the time-integral of the detector plane intensity, detector signals, and the class scores for the time-lapse image classification network. While the time-static network misclassifies the object for an "automobile," the time-lapse image classification diffractive network correctly predicts the object to be a "ship" (predicted label: 8). We also show in Figure 3c the confusion matrices calculated over 10 000 test images of the CIFAR-10 dataset: the time-lapse image classification diffractive network performs consistently better than the time-static one for all the CIFAR-10 data classes.

Note also that the time-lapse image classification diffractive network designed with nontrainable detector exponents (i.e., $\gamma_{c,\pm} = 1$) achieved a blind testing accuracy of 60.35% on the same grayscale CIFAR-10 test dataset (see Figure S1, Supporting Information), performing much better than the time-static one for all the CIFAR-10 data classes. The diffractive layers for all these networks are shown in Figure S2, Supporting Information.

During the training of the time-lapse diffractive networks, we followed a method similar to the "dropout" method, which is used in deep learning to reduce overfitting and improve the generalization of a trained model.[37] We defined a hyperparameter p which is the probability that a point on the object-plane grid is "active" during training, i.e., the probability that the object is positioned at that lateral point during the signal integration at the detector. All the time-lapse networks described thus far were trained with $p = 0.5$. As we describe below, the resilience of the trained time-lapse image classification diffractive networks to deviations from the training settings can be improved by a proper choice of p, which is intuitively similar to the dropout strategy in deep learning literature.

Related to this hyperparameter p, next, we explored the impact of decreasing the number of lateral shifts, N, on the blind testing accuracy of time-lapse classifiers (see **Figure 4**). The value for each data point in Figure 4 represents the mean of the classification accuracies over 25 independent blind tests with the same N. For Figure 4a, these N lateral displacements were restricted to coincide with the predetermined training grid points, and for the case of $N < m^2$, $m^2 - N$ of the m^2 lateral shifts were randomly eliminated (not used). For Figure 4b, however, the N lateral displacements were randomly selected without following the training grid points. As we can see in Figure 4a, the blind testing accuracy decreases as N is decreased; however, the slope of this performance degradation varies depending on the training hyperparameter p. For example, in the case of the time-lapse image classification diffractive network shown in Figure 3b, trained with $p = 0.5$ (green curve in Figure 4a), the test accuracy drops from 62.03% to 60.69% and 59.37% as N decreases from 25 to 15 and 10, respectively. Compare this with the case of a time-lapse diffractive network trained with $p = 1.0$ (red curve in Figure 4a), for which the classification accuracy is affected much more severely and decreases from 61.61% to 59.61% and 57.45% as N is decreased from 25 to 15 and 10, respectively. We see that networks trained with lower p-values show less

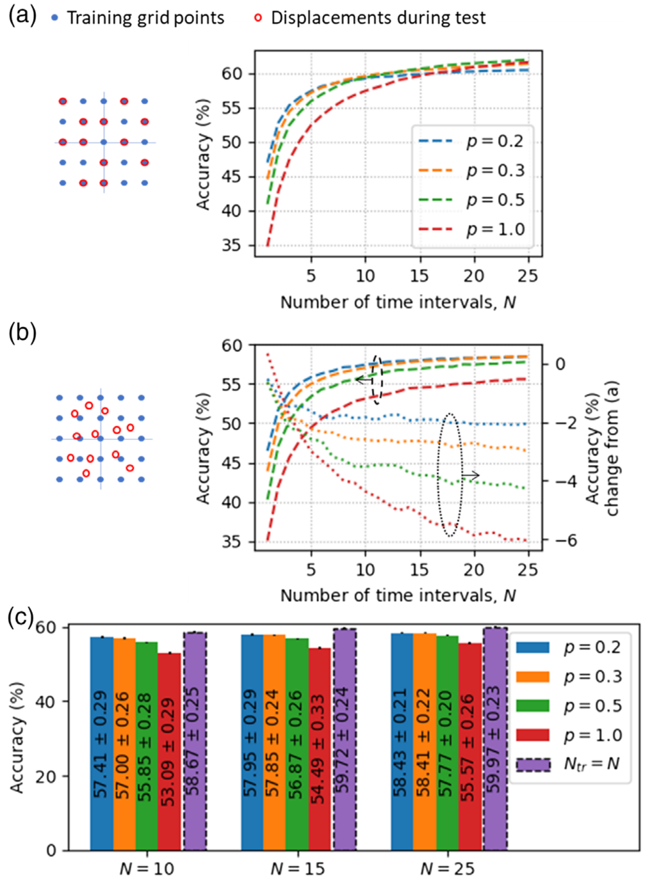

Figure 4. Impact of the lateral shifts. a) Impact of decreasing the number of lateral shifts (N) on the blind testing accuracy while confining the input object displacements to the training grid points; the time-lapse D^{2}NNs were trained with different p-values. Diffractive networks trained with lower p-values are less affected by a decrease in N. b) Impact of random off-grid lateral displacements of input objects during the blind testing of time-lapse diffractive networks trained with displacements confined to a predefined grid. Time-lapse diffractive classifiers trained with lower p-values are more resilient to such deviation from the training settings. c) Improvement of blind testing accuracies for a given N, by training the time-lapse diffractive network with $N_{tr} = N$ arbitrary/random displacements within the range $2s_{\max} \times 2s_{\max}$ instead of training with a set of fixed lateral displacements defined by a predetermined lateral grid. For (a)–(c), the values (errors) corresponding to the data points represent the mean (standard deviation) values calculated through the blind testing of the same trained network 25 times, every time with N arbitrary lateral displacements of the input objects.

sensitivity to decreasing N, which is further corroborated by the curves corresponding to two other time-lapse diffractive networks trained with $p=0.2$ and $p=0.3$.

Another advantage of training with lower p-values is decreased sensitivity to the exact object positions (see Figure 4b). For Figure 4b, we selected the N lateral displacements without following the training grid points, allowing the object to be displaced (during the time-lapse imaging process) to N arbitrary, randomly selected points within the area $2s_{\max} \times 2s_{\max}$. In general, for a given N, the blind testing accuracies corresponding to such arbitrary displacements (left y-axis of Figure 4b) are lower than their counterparts for the on-grid displacements shown in Figure 4a. However, the degradation in classification accuracy, which is shown on the right y-axis of Figure 4b, is much smaller when p is lower. For example, at $N=25$, the mean accuracy drop is ≈2% for the diffractive network trained with $p=0.2$, whereas the accuracy drop is ≈6% for the $p=1.0$ diffractive network.

The accuracy of time-lapse diffractive network-based image classifiers for arbitrary lateral displacements of the input objects can be improved by utilizing such random displacements of the objects during the training, rather than training with a predetermined grid of lateral displacements. For this, the training hyperparameters p and m can be absorbed into a single hyperparameter N_{tr}, where N_{tr} refers to the number of arbitrary displacements within $2s_{\max} \times 2s_{\max}$. To demonstrate this, we trained three time-lapse diffractive networks with $N_{\text{tr}}=10$, $N_{\text{tr}}=15$, and $N_{\text{tr}}=25$ and compared their accuracies for $N=10$, $N=15$, and $N=25$ arbitrary displacements of the input objects, respectively, against the classification accuracies of the time-lapse diffractive networks reported in Figure 4a,b. The result of this comparison is shown in Figure 4c: for $N=10$, $N=15$, and $N=25$ arbitrary lateral displacements during the time-lapse imaging process, the mean blind testing accuracies of the corresponding $N_{\text{tr}}=N$ diffractive networks are 1.26%, 1.77%, and 1.54%, respectively, higher than the accuracies of the $p=0.2$ time-lapse diffractive network. This generalization improvement and the inference accuracy increase are due to using arbitrary random lateral displacements of the input objects during the training process instead of blindly applying such random lateral shifts only during the testing phase.

3. Discussion

In previous work, we reported a significant improvement in diffractive network inference performance by ensemble learning and combining the output of several different diffractive networks. For example, mean blind testing accuracies of 61.14% and 62.13% on the CIFAR-10 test set were reported for ensembles of 14 and 30 different D^{2}NNs, respectively.[32] However, the improvement with such a strategy is accompanied by a sacrifice in the compactness of the optical hardware and increased complexity in aligning several diffractive networks within the ensemble. Another shortcoming of ensemble learning of diffractive networks is the large training time. In our previous work, 1252 diffractive models were trained, and ensemble pruning was then performed to arrive at the final design.[32] Time-lapse diffractive network-based image classification provides blind testing accuracies comparable to ensemble learning with only a single trained diffractive network. For comparison, the time-lapse diffractive network of Figure 3b gives 62.03% blind testing accuracy on CIFAR-10 test images. The trade-off for such an advantage is the increase in the imaging/classification time due to the lateral shifts of the objects. However, the alignment and synchronization requirements associated with diffractive network ensembles are evaded. Also, the training of a time-lapse diffractive classifier takes ≈20 h on an NVIDIA GeForce RTX 3090 GPU (see the Experimental Section), which is orders of magnitude less than the time required to design an ensemble of diffractive networks working together.

Regarding the implementation of time-lapse diffractive network-based image classification, spatial light modulators (SLMs) can be used to perform the lateral displacements of the input objects digitally if a digital representation of each object is available. In an alternative implementation, the diffractive layers and the detectors could be mounted on a movable stage to shift the entire system with respect to the object or input FOV. Perhaps, the simplest implementation of time-lapse diffractive network-based image classification would exploit the natural jitter or movement of the input objects during the integration time of the class detectors. As shown in Figure 4c, ≈60% blind testing accuracy on CIFAR-10 test images can be reached with arbitrary object displacements during the time-lapse inference.

A point worth emphasizing is that although the displacements in Figure 4b (as well as in the other figures) are shown to be discrete and mostly disconnected to reflect our numerical analysis, the proposed scheme is also applicable for the continuous, connected motion of objects as it occurs in real life during time-lapse imaging. As long as the motion of the object covers the training lateral range $2s_{\max} \times 2s_{\max}$ within the detector integration time, we expect a performance improvement due to time-lapse inference over time-static inference. We corroborate this fact with additional numerical analysis, the results of which are presented in Figure S10, Supporting Information. In the simulations shown in Figure S10a,b, Supporting Information, we tested the already trained time-lapse diffractive network of Figure 4c for different continuous object trajectories (with connected sets of random points), irregularly spanning an area of $2s_{\max} \times 2s_{\max}$—matching the training scheme in terms of the maximum lateral displacements. For these irregular, continuous trajectories that fit within an area of $\approx 2s_{\max} \times 2s_{\max}$ on the input plane, the average blind testing accuracy over 10 000 CIFAR-10 test images was 60.68% and 61.00% for Figure S10a,b, Supporting Information, respectively. In contrast, we obtained a decreased blind testing accuracy of 56.18% for the larger continuous/connected trajectory of Figure S10c, Supporting Information, that spans an area larger than $2s_{\max} \times 2s_{\max}$; this reduced inference accuracy is due to the object displacements that lie outside the region for which the network was trained (green dashed square in Figure S10, Supporting Information). The accuracy becomes 61% when only the portion of this trajectory within the $2s_{\max} \times 2s_{\max}$ area is used, as shown in Figure S10b, Supporting Information.

One of the challenges regarding the practical implementations of our proposed framework is the realization of the relative displacements between the objects and the input aperture of the diffractive network. The tolerance of the time-lapse diffractive network to imprecise displacements as well as deviations from the training conditions (such as the one shown in Figure S10c, Supporting Information) can be improved by using "vaccination" schemes,[38] where such deviations are modeled as random noise during the training phase to make the diffractive network designs robust to them. However, as discussed earlier in relation to the results of Figure 4c and S10, Supporting Information, the proposed time-lapse inference scheme provides a significant improvement over time-static classification even for arbitrary/random displacements of the input objects (regardless of being continuous/connected or disconnected). The sensitivity of a time-lapse diffractive network to these random displacements between the object and the diffractive network input aperture can become an issue if the input objects move significantly outside of the lateral shift area used during the training, as shown in Figure S10c, Supporting Information. This can potentially be addressed by further increasing $s_{\max}$ in the training phase. Other practical concerns, such as fabrication errors and mechanical misalignments between the diffractive layers, could also be addressed similarly using a form of vaccination[38] during the training phase, which guides the evolution of the diffractive network layers to be resilient against such random imperfections; in general, a downside of the vaccination strategy is a reduction in the independent degrees of freedom available within the diffractive network layers, which can be mitigated by, e.g., increasing the width of each diffractive layer or using a deeper network architecture with more layers. To overcome the detection noise associated with optoelectronic detectors at the output plane, the SNR of the time-lapse diffractive network can be improved by penalizing low output diffraction efficiencies during the training phase, which will be detailed next.

In terms of the photon budget, the output diffraction efficiency of the time-lapse diffractive image classifier of Figure 3b is 0.6%, which refers to the ratio of the total incident energy on the output detectors with respect to the total energy at the input aperture. The output diffraction efficiency of a diffractive network can be significantly improved by introducing an additional loss term that penalizes low efficiency designs. This diffraction efficiency-related term can be added to the total training loss function with a linear weight (η), which is a hyperparameter that can be tuned depending on the acceptable trade-off between the inference accuracy and the output diffraction efficiency (see the Experimental Section). To illustrate this trade-off, we trained two new time-lapse diffractive networks with $\eta = 0.85$ and $\eta = 0.9$, where the former resulted in a blind testing accuracy of 60.66% (CIFAR-10) and an output diffraction efficiency of 3.2%, whereas the latter, $\eta = 0.9$ design, resulted in a blind testing accuracy of 58.02% and an output diffraction efficiency of 32.7%.

We should also note that the inference accuracy of the time-lapse diffractive network of Figure 3b is nonuniform across different classes, e.g., varying from 40.5% for class 3 to 75.3% for class 1. This inference nonuniformity from class to class can be reduced by using class-specific loss weights during the training phase. To demonstrate this, we calculated the mean loss (L_c) of the diffractive network of Figure 3b for all the classes $c = 0, 1, \cdots, 9$. Then, we trained two new diffractive networks: one with $w_c \propto L_c$ and the other with $w_c \propto \exp(2L_c)$, where w_c refers to the class-specific weights used to calculate the training loss function. We show these class-specific loss weights and the resulting confusion matrices for these time-lapse diffractive networks in Figure S3, Supporting Information. For the time-lapse diffractive network trained with $w_c \propto L_c$, the average blind testing accuracy is 61.36% (for CIFAR-10) and the individual class accuracies range from 42.9% to 72.5%; on the other hand, for the time-lapse diffractive network trained with $w_c \propto \exp(2L_c)$, the average blind testing accuracy is 60.26% and the individual class accuracies range from 50.3% to 67.4%, see Figure S3c, Supporting Information. These results indicate that the statistical inference nonuniformity from class to class can be reduced, if desired, by using class-specific loss weights during the training phase with a small trade-off in the average classification accuracy of the time-lapse diffractive network.

Although a time-lapse diffractive network has access to enriched information regarding an input sample due to lateral shifts of its field-of-view, it still needs a rich training dataset for a successful design and generalization. Stated differently, time-lapse diffractive network training with a certain number (N)of lateral displacements of the objects cannot compensate for a reduction of the training dataset size because the unique spatial features of each image of the training dataset help the diffractive networks' generalization. For example, training a time-lapse diffractive network with $m = 2$ (i.e., $N = 4$ lateral displacements) using a reduced dataset that has fourfold smaller number of examples achieved a lower blind testing accuracy of 51.78% for CIFAR-10. As another example, training a time-lapse diffractive network with $m = 5$ (i.e., $N = 25$ lateral displacements) using a reduced dataset that has 25-fold smaller number of examples achieved an even lower blind testing accuracy of 44.62%. These indicate that the generalization improvement of time-lapse diffractive networks needs diverse training datasets and time-lapse imaging alone cannot overcome poor generalization due to reductions in the size of the training dataset.

The use of a pair of detectors for each data class in a differential detection scheme was shown to improve the inference performance of time-static diffractive networks.[11] This differential detection scheme helps to virtually circumvent the nonnegativity constraint on the class scores, imposed by optoelectronic detectors which are sensitive only to the intensity of light. This inference performance improvement due to the differential detection scheme is also translated to time-lapse diffractive networks. For example, we trained, from scratch, a time-lapse diffractive network with the same hyperparameters as those of the network shown in Figure 3b, but this time used the sum of the optical signals from the two detectors assigned to each class as the class score, instead of using their difference signal. The blind testing accuracy with this nondifferential detection scheme was 61.02% compared to the 62.03% blind testing accuracy of the differential scheme.

While time-lapse image classification significantly boosts the inference accuracy of a single D^2NN on the classification of complex objects, there remains plenty of room for improvement to potentially close the large performance gap with their electronic counterparts, convolutional deep neural networks.[32] One possible avenue for such an improvement could be the incorporation of ensemble learning with time-lapse image classification, where the outputs of diversely trained time-lapse D^2NNs could be combined for further improvement in generalization and statistical inference. To numerically demonstrate that ensemble learning can improve the inference of time-lapse networks, we trained

20 different time-lapse image classifiers, all with $m = 5$ and $s_{\max} = 5.33\lambda$. To diversify the individual diffractive classifiers that are trained, we used two different strategies[32] as illustrated in Figure S4, Supporting Information; first, we used different encodings, i.e., amplitude encoding or phase encoding of the object information at the input plane; second, we used fixed (nontrainable) amplitude-only filters with different spatial profiles, either immediately following the object or at the Fourier plane of a 4*f* system following the object.[32] The trained diffractive layers for these 20 time-lapse image classification networks are shown in Figure S5, Supporting Information. The blind testing accuracies of these individual time-lapse diffractive networks range from 48.18% to 60.79%. In Figure S6–S9, Supporting Information, we present four ensembles with constituents selected from these 20 individual time-lapse diffractive networks (see the Supplementary Information for the ensemble design strategy). For example, Figure S6, Supporting Information, shows an ensemble of $M = 20$ time-lapse diffractive networks that attain 65.05% blind testing accuracy on CIFAR-10 test images utilizing $N = 25$ lateral object displacements (see the Supplementary Information for details); as another example, a smaller ensemble with $M = 10$ time-lapse diffractive networks attained 64.21% blind testing accuracy utilizing $N = 11$ lateral displacements (Figure S8, Supporting Information). Two other ensembles, with $M = 20$ ($N = 25$) and $M = 10$ ($N = 8$), are also reported in Figure S7 and S9, Supporting Information, respectively, with similar performances. Further optimization of the ensemble hyperparameters and the encoding/filtering strategies of the input object information could further improve the inference accuracy of time-lapse diffractive network ensembles.

Moreover, in the same way that the time-lapse scheme utilizes the complementary information resulting from the input objects that are laterally shifted, other attributes of light such as polarization or wavelength could also be utilized to further improve the inference performance of time-lapse diffractive networks.[9,39] For example, time-lapse diffractive networks can be trained to work with RGB images instead of grayscale images to benefit from the complementary information carried by different color channels. The incorporation of optical nonlinearities between the diffractive layers of D^{2}NNs could also extend their approximation capability and consequently improve their statistical inference; for further details, see the Supplementary Information of ref. [6], where the impact of optical nonlinearities within a D^{2}NN architecture was discussed for the first time. All of these constitute possible future directions to explore for further decreasing the performance gap between electronic deep neural networks and D^{2}NNs.

In summary, we reported a time-lapse diffractive network-based image classification scheme for significantly improving the performance of D^{2}NN classifiers with only a single trained diffractive network. The presented time-lapse diffractive network scheme could be vital for realizing compact, low-cost, and passive optical processors for all-optical spatiotemporal analysis of information.

4. Experimental Section

Forward Model: The propagation of coherent light across K+2 parallel planes defined by the input (object) plane, K successive diffractive layers, and the output (detector) plane is modeled using the Rayleigh–Sommerfeld theory of scalar diffraction,[40] according to which the propagation of a complex wave $U(x, y)$ through a distance z in free-space is described by a linear shift-invariant system with an impulse response defined as follows

$$h(x, y; z) = \frac{z}{r^2}\left(\frac{1}{2\pi r} + \frac{1}{j\lambda}\right)\exp\left(j\frac{2\pi r}{\lambda}\right) \tag{1}$$

where λ is the illumination wavelength, $r = \sqrt{x^2 + y^2 + z^2}$, and $j = \sqrt{-1}$. Upon propagation through the free-space separating layer l—1 and layer l, the complex field is modulated by the spatially varying complex transmittance $t_l(x, y)$ of layer l, i.e.

$$\begin{aligned} U(x, y; z_l) = t_l(x, y) \iint & h(x - x', y - y';\ z_l - z_{l-1}) \\ & \times U(x', y';\ z_{l-1})dx'dy' \end{aligned} \tag{2}$$

$$t_l(x, y) = a_l(x, y)\exp(j\varphi_l(x, y)) \tag{3}$$

Here, z_l is the axial coordinate of the lth plane, and $l = 1, \cdots, K$, whereas $a_l(x, y)$ and $\varphi_l(x, y)$ are the amplitude and the phase of the complex field transmittance $t_l(x, y)$. For the phase-only diffractive networks reported in this work, $a_l(x, y)$ is assumed to be 1.

In a differential classification scheme, each of the ten classes of the CIFAR-10 dataset is assigned to two detectors: a virtual positive detector and a virtual negative detector. $D_{c,+}$ ($D_{c,-}$) denotes the active area of the positive (negative) detector assigned to class c, $c = 0, 1, \cdots, 9$. For the time-static diffractive networks, the detector signals $I_{c,\pm}$, based on which the class scores are computed, are proportional to the detector powers $P_{c,\pm}$, where

$$P_{c,\pm} = \iint_{D_{c,\pm}} |U(x, y; z_{K+1})|^2 \mathrm{d}x\mathrm{d}y \tag{4}$$

For the time-lapse diffractive network, light is assumed to be integrated at the detectors over N intervals of δ_t duration each. The object function (O) during the nth interval can be expressed as

$$O_n(x, y) = O(x - x_n, y - y_n) \tag{5}$$

$$n = 1, \cdots, N \tag{6}$$

$$(x_n, y_n) \in X \times Y \tag{7}$$

$$X = Y = \mathrm{linspace}(-s_{\max}, s_{\max}, m) \tag{8}$$

where $\mathrm{linspace}(-s_{\max}, s_{\max}, m)$ denotes the set of m linearly spaced values between $-s_{\max}$ and $s_{\max}$. Accordingly, the optoelectronic signals at the detectors are proportional to the integrated photon signals

$$E_{c,\pm} = \alpha \int_0^{N\delta_t} P_{D_{c,\pm}}(t)\mathrm{d}t \tag{9}$$

Here, α is an optoelectronic detector-specific constant, and we assume that the propagation delay of light between the object plane and the detector plane is negligible compared to δ_t.

The detectors are assigned the exponents $\gamma_{c,\pm}$, which operate on the optoelectronic signals $E_{c,\pm}$ (after $E_{c,\pm}$ are normalized to have a maximum value of 1) and generate the detector signals $I_{c,\pm}$

$$I_{c,\pm} = (E_{c,\pm})^{\gamma_{c,\pm}} \tag{10}$$

Finally, the differential class scores are calculated as

$$z_c = \frac{I_{c,+} - I_{c,-}}{I_{c,+} + I_{c,-}} \tag{11}$$

and the prediction for the object class is defined to be $\arg\max_c z_c$.

Numerical Implementation: When numerically modeling light propagation through the diffractive networks, the grid spacing along the transverse directions (x and y) was chosen to be $\approx 0.53\lambda$. The Rayleigh–Sommerfeld convolution integrals were computed using the angular spectrum method[40] based on the fast Fourier transform (FFT). For all the results presented in this article, the diffractive networks consisted of five phase-only diffractive layers, axially separated by 40λ. Each layer comprised 200×200 diffractive features/neurons, the phases of which were trainable. The (physical) size of each diffractive neuron was assumed to be $\approx 0.53\lambda \times 0.53\lambda$.

The RGB images in the CIFAR-10 dataset were converted to grayscale to represent the input objects illuminated by a monochromatic and spatially coherent wave. The objects were resized to span an area of $44.8\lambda \times 44.8\lambda$. The object information was assumed to be encoded in the phase channel of the input light, i.e., within the input field of view

$$U(x, y; z_0) = \exp(j2\pi O(x, y)) \tag{12}$$

where $O(x, y)$ is the object function, with its values normalized to lie between 0 and 1. On the output plane, the active area of each detector was assumed to be $6.4\lambda \times 6.4\lambda$, and the spacing between the detectors was $\approx 4.27\lambda$ along both x and y directions (see Figure 1).

Training: The diffractive networks were trained using the cross-entropy loss function together with a loss term penalizing low diffraction efficiency at the output detectors. The differential class scores $\{z_c\}_{c=0}^{9}$ were converted to probabilities $\{q_c\}_{c=0}^{9}$ over the data classes using the softmax function, i.e.

$$q_c = \frac{\exp(\beta z_c)}{\sum_i \exp(\beta z_i)} \tag{13}$$

where $\beta = 10$ was used. The training loss function was defined as

$$\mathscr{L} = \mathscr{L}_{\mathrm{CE}} + \eta \mathscr{L}_{\mathrm{eff}} \tag{14}$$

$$\mathscr{L}_{\mathrm{CE}} = -\sum_{c=0}^{9} \delta_{ck} \log q_c \tag{15}$$

$$\mathscr{L}_{\mathrm{eff}} = 1 - \frac{E_{\mathrm{D}}}{E_{\mathrm{illum}}} \tag{16}$$

where $\mathscr{L}_{\mathrm{CE}}$ is the cross-entropy loss, k is the (true) label, and δ_{ck} is the Kronecker delta function, i.e., $\delta_{ck} = 1$ if $c = k$ and 0 otherwise. $\mathscr{L}_{\mathrm{eff}}$ is introduced to penalize low diffraction efficiency at the output detectors, where $E_{\mathrm{D}} = \sum_{c=0}^{9}(E_{c,+} + E_{c,-})$ is the total incident energy on the 20 detectors and E_{illum} is the total energy of the illumination beam at the input aperture. η is a training hyperparameter, which is used to fine-tune the output diffraction efficiency of the resulting diffractive networks.

The trainable parameters of the model were trained by minimizing the loss $\mathscr{L}$ using the adaptive momentum ("Adam") stochastic gradient descent algorithm.[41] The forward model was implemented using the open-source deep learning library TensorFlow.[42] The automatic differentiation functionality of TensorFlow was exploited to facilitate the gradient computations for optimization. A batch size of 8 was used to implement the stochastic gradient descent. The built-in TensorFlow implementation of Adam optimizer was used with the default values except for the learning rate, which had an initial value of 0.001 and was reduced by a factor of 0.7 every eight epochs.

All the networks were trained for 100 epochs using 45 000 images from the training set of the CIFAR-10 dataset. The remaining 5000 images of the CIFAR-10 training set were left out for validation, i.e., after every epoch, the accuracy of the model on these 5000 images was evaluated. The model state at the end of the epoch for which the validation accuracy was maximum was ultimately used for blind testing.

The training time of the time-lapse diffractive networks depended upon the hyperparameters m and p. For $m = 5$ and $p = 0.5$, the training took $\approx$20 h on an NVIDIA GeForce RTX 3090 GPU in a machine running on Windows 10.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The Ozcan Research Group at UCLA acknowledges the support of U.S. Department of Energy (DOE), Office of Basic Energy Sciences, Division of Materials Sciences and Engineering under Award # DE-SC0023088.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

M.S.S.R. implemented and trained the diffractive network models and performed the numerical simulations for preparing the results reported in this work; M.S.S.R. and A.O. analyzed the results and wrote the manuscript; A.O. initiated and supervised the project.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

diffractive deep neural networks ($\mathrm{D^2NN}$), diffractive computing, diffractive deep neural networks, diffractive networks, optical computing

Received: November 4, 2022
Revised: November 25, 2022
Published online: February 21, 2023

[1] M. I. Jordan, T. M. Mitchell, *Science* **2015**, *349*, 255.
[2] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[3] W. Knight, https://www.wired.com/story/ai-great-things-burn-planet/ (accessed: August 2022).
[4] R. Toews, https://www.forbes.com/sites/robtoews/2020/06/17/deep-learnings-climate-change-problem/ (accessed: August 2022).
[5] R. Hamerly, *IEEE Spectr.* **2021**, *58*, 30.
[6] X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, A. Ozcan, *Science* **2018**, *361*, 1004.
[7] O. Kulce, D. Mengu, Y. Rivenson, A. Ozcan, *Light Sci. Appl.* **2021**, *10*, 25.
[8] O. Kulce, D. Mengu, Y. Rivenson, A. Ozcan, *Light Sci. Appl.* **2021**, *10*, 196.
[9] J. Li, Y.-C. Hung, O. Kulce, D. Mengu, A. Ozcan, *Light Sci. Appl.* **2022**, *11*, 153.
[10] D. Mengu, Y. Luo, Y. Rivenson, A. Ozcan, *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 1.
[11] J. Li, D. Mengu, Y. Luo, Y. Rivenson, A. Ozcan, *Adv. Photonics* **2019**, *1*, 046001.
[12] Y. Luo, D. Mengu, N. T. Yardimci, Y. Rivenson, M. Veli, M. Jarrahi, A. Ozcan, *Light Sci. Appl.* **2019**, *8*, 112.
[13] M. Veli, D. Mengu, N. T. Yardimci, Y. Luo, J. Li, Y. Rivenson, M. Jarrahi, A. Ozcan, *Nat. Commun.* **2021**, *12*, 37.

[14] Y. Luo, Y. Zhao, J. Li, E. Çetintaş, Y. Rivenson, M. Jarrahi, A. Ozcan, *eLight* **2022**, *2*, 4.
[15] M. S. S. Rahman, A. Ozcan, *ACS Photonics* **2021**, *8*, 3375.
[16] D. Mengu, A. Ozcan, *Adv. Opt. Mater.* **2022**, *10*, 2200281.
[17] B. Bai, Y. Luo, T. Gan, J. Hu, Y. Li, Y. Zhao, D. Mengu, M. Jarrahi, A. Ozcan, *eLight* **2022**, *2*, 14.
[18] C. Isil, D. Mengu, Y. Zhao, A. Tabassum, J. Li, Y. Luo, M. Jarrahi, A. Ozcan, *Sci. Adv.* **2022**, *8*, 48.
[19] Y. Luo, D. Mengu, A. Ozcan, *Sci. Rep.* **2022**, *12*, 7121.
[20] C. Qian, X. Lin, X. Lin, J. Xu, Y. Sun, E. Li, B. Zhang, H. Chen, *Light Sci. Appl.* **2020**, *9*, 59.
[21] Z. Huang, P. Wang, J. Liu, W. Xiong, Y. He, J. Xiao, H. Ye, Y. Li, S. Chen, D. Fan, *Phys. Rev. Appl.* **2021**, *15*, 014037.
[22] Z. Huang, Y. He, P. Wang, W. Xiong, H. Wu, J. Liu, H. Ye, Y. Li, D. Fan, S. Chen, *Opt. Express* **2022**, *30*, 5569.
[23] J. Zhang, Z. Ye, J. Yin, L. Lang, S. Jiao, *Opt. Express* **2022**, *30*, 26728.
[24] P. Buske, A. Völl, M. Eisebitt, J. Stollenwerk, C. Holly, *Opt. Express* **2022**, *30*, 22798.
[25] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li, T. J. Cui, *Nat. Electron.* **2022**, *5*, 113.
[26] T. Zhou, X. Lin, J. Wu, Y. Chen, H. Xie, Y. Li, J. Fan, H. Wu, L. Fang, Q. Dai, *Nat. Photonics* **2021**, *15*, 367.
[27] T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, Q. Dai, *Phys. Rev. Lett.* **2019**, *123*, 023901.
[28] X. Luo, Y. Hu, X. Ou, X. Li, J. Lai, N. Liu, X. Cheng, A. Pan, H. Duan, *Light Sci. Appl.* **2022**, *11*, 158.
[29] J. Shi, J. Shi, Y. Chen, X. Zhang, X. Zhang, *Opt. Lett.* **2022**, *47*, 605.
[30] J. Shi, L. Zhou, T. Liu, C. Hu, K. Liu, J. Luo, H. Wang, C. Xie, X. Zhang, *Opt. Lett.* **2021**, *46*, 3388.
[31] CIFAR-10 and CIFAR-100 datasets, , ≈https://www.cs.toronto.edu/≈kriz/cifar.html (accessed: August 2022).
[32] M. S. S. Rahman, J. Li, D. Mengu, Y. Rivenson, A. Ozcan, *Light Sci. Appl.* **2021**, *10*, 14.
[33] W. Bishara, T.-W. Su, A. F. Coskun, A. Ozcan, *Opt. Express* **2010**, *18*, 11181.
[34] A. Greenbaum, W. Luo, B. Khademhosseinieh, T. W. Su, A. F. Coskun, A. Ozcan, *Sci. Rep.* **2013**, *3*, 1717.
[35] A. Greenbaum, A. Feizi, N. Akbari, A. Ozcan, *Opt. Express* **2013**, *21*, 12469.
[36] R. Dahl, M. Norouzi, J. Shlens, in *2017 IEEE Int. Conf. Computer Vision (ICCV)*, IEEE, Piscataway, NJ **2017**, pp. 5439–5448.
[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov *J. Mach. Learn. Res.* **2014**, *15*, 1929.
[38] D. Mengu, Y. Zhao, N. T. Yardimci, Y. Rivenson, M. Jarrahi, A. Ozcan, *Nanophotonics* **2020**, *9*, 4207.
[39] W. Luo, Y. Zhang, A. Feizi, Z. Göröcs, A. Ozcan, *Light Sci. Appl.* **2016**, *5*, e16060.
[40] J. W. Goodman, *Introduction to Fourier Optics*, Roberts and Company Publishers, Englewood, CO, USA **2005**.
[41] D. P. Kingma, J. Ba, (*Preprint*) *arXiv*: 1412.6980, v9, submitted: Jan **2017**.
[42] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, in *Proc. 12th USENIX Conf. Operating Systems Design and Implementation*, USENIX Association, Savannah, GA, USA **2016**, pp. 265–283.