Research Article Vol. 5, No. 7 / July 2018 / Optica 803

Imaging through glass diffusers using densely connected convolutional networks

Shuai Li,[1,*] Mo Deng,[2] Justin Lee,[3] Ayan Sinha,[1,5] and George Barbastathis[1,4]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
[3]Institute for Medical Engineering Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
[4]Singapore-MIT Alliance for Research and Technology (SMART) Centre, One Create Way, Singapore 117543, Singapore
[5]Current address: Magic Leap Inc, 1376 Bordeaux Drive, Sunnyvale, California 94089, USA
*Corresponding author: shuaili@mit.edu

Received 15 May 2018; revised 17 June 2018; accepted 17 June 2018 (Doc. ID 331810); published 6 July 2018

Computational imaging through scatter generally is accomplished by first characterizing the scattering medium so that its forward operator is obtained and then imposing additional priors in the form of regularizers on the reconstruction functional to improve the condition of the originally ill-posed inverse problem. In the functional, the forward operator and regularizer must be entered explicitly or parametrically (e.g., scattering matrices and dictionaries, respectively). However, the process of determining these representations is often incomplete, prone to errors, or infeasible. Recently, deep learning architectures have been proposed to instead learn both the forward operator and regularizer through examples. Here, we propose for the first time, to our knowledge, a convolutional neural network architecture called "IDiffNet" for the problem of imaging through diffuse media and demonstrate that IDiffNet has superior generalization capability through extensive tests with well-calibrated diffusers. We also introduce the negative Pearson correlation coefficient (NPCC) loss function for neural net training and show that the NPCC is more appropriate for spatially sparse objects and strong scattering conditions. Our results show that the convolutional architecture is robust to the choice of prior, as demonstrated by the use of multiple training and testing object databases, and capable of achieving higher space–bandwidth product reconstructions than previously reported.

OCIS codes: (100.3190) Inverse problems; (100.4996) Pattern recognition, neural networks; (110.1758) Computational imaging.

https://doi.org/10.1364/OPTICA.5.000803

1. INTRODUCTION

Imaging through random media [1,2] remains one of the most useful as well as challenging topics in computational optics. The reason is that scattering impedes information extraction from the wavefront in two distinct, albeit related ways. First, light scattered at angles outside the system's numerical aperture is lost; second, the relative phases among spatial frequencies that pass are scrambled—convolved with the diffuser's own response. In most cases, the random medium is not known, or it is unaffordable to characterize it completely. Even if the random medium and, hence, the convolution kernel are known entirely, deconvolution is highly ill-posed and prone to noise-induced artifacts.

Therefore, the strategy to recover the information, to the degree possible, must be two-pronged: first, to characterize the medium as well as possible so that at least errors in the deconvolution due to incomplete knowledge of the medium's response may be mitigated; second, to exploit additional *a priori* knowledge about the class of objects being imaged so that the inverse problem's solution space is reduced and spurious solutions are excluded. These two strategies are summarized by the well-known Tikhonov–Wiener optimization functional for solving inverse problems as

$$\hat{f} = \text{argmin}_f\{\|g - Hf\|^2 + \alpha\Phi(f)\}, \tag{1}$$

where H is the forward operator in the optical system $g = Hf$, f is the unknown object, g is the raw intensity image (or images if some form of scanning is involved), $\Phi(\cdot)$ is the regularizer function, α is the regularization parameter controlling the relative contribution of the two terms in the functional, and $\hat{f}$ is the estimate of the object.

The forward operator H includes the effects of the scatterer, as well as of the optical system utilized in any situation. A number of ingenious strategies have been devised to design forward operators that improve the imaging problem's condition, most famously by using nonlinear optics [3,4] or stimulated emission [5]. Restricting oneself to linear optics, structured illumination [6–8] is an effective strategy that modulates object information onto better-behaved spatial frequencies.

Several approaches characterize the random medium efficiently. One method is to measure the transmission matrix (TM) of the medium by interferometry or wavefront sensing [9–11]. Alternatively, one may utilize the angular memory effect in speckle correlation [12–16]. The angular memory principle states that rotating the incident beam over small angles does not change the resulting speckle pattern but only translates it over a small distance [17,18]. In this case, computing the autocorrelation of the output intensity and deconvolving it by the speckles' autocorrelation function, which is a sharply peaked function [19], result in the autocorrelation of the input field. Then, the object is recovered from its own autocorrelation using the Gerchberg–Saxton–Fienup (GSF) algorithm [20,21] with additional prior constraints.

The regularizer Φ expresses prior knowledge by penalizing unacceptable objects so the optimization is prohibited from landing onto them; alternatively, the priors expressed by the regularizer can be thought of as helping to resolve nonuniqueness due to the ill-posed nature of the forward operator. In the case of strong scattering, it is common to say that information "is lost" because it is convolved into the high spatial frequencies escaping the system aperture. (The opposite may also be possible: a cleverly designed scattering medium may bring high-spatial frequency information back into the aperture, by convolving it to low spatial frequencies [6,22,23].) However, the prior may help to recover the missing information by enforcing properties such as edge sharpness or, more generally, sparsity, positivity, etc. During the past two decades, owing to efforts by Grenander [24], Candés, *et al.* [25], and Brady *et al.* [26], the use of sparsity priors was popularized and proved to be effective in a number of contexts including random media. For example, Liu *et al.* successfully recovered the 3D positions of multiple LEDs embedded in turbid scattering media by taking phase-space measurements and imposing the $\mathcal{L}1$ sparsity prior [27].

Instead of establishing H and Φ independently and explicitly from measurements and prior knowledge, an alternative approach is to *learn* both operators simultaneously through examples of objects imaged through the random medium. To our knowledge, the first instance of this strategy was by Horisaki *et al.* [28]. In that paper, a support vector regression (SVR) learning architecture was used to learn the scatterer and the prior of faces being imaged through. The approach was effective in that the SVR learned correctly to reconstruct face objects; it also elucidated the generalization limitations of SVRs, which are shallow fully connected two-layer architectures: for example, when presented with nonface objects the SVR would still respond with one of its learned faces as a reconstruction. A deeper fully connected architecture in the same learning scheme has been proposed recently [29]. The Horisaki paper was the first, to our knowledge, to use machine learning in the computational imaging context; it certainly influenced our own work on lensless imaging [30] and other related works [31–35].

In this paper, we propose for the first time, to our knowledge, two innovations in the use of machine learning for imaging through scatter: the first is the use of the convolutional neural network (CNN) architecture, [36] and the second is the use of a negative Pearson correlation coefficient (NPCC) as the loss function. Different from fully connected network architectures, in the CNN each neuron is connected to only a few nearby neurons in the previous layer, and the same set of weights is used for every neuron. The fewer number of connections and weights reduces the complexity of the CNN architecture and makes convolutional layers relatively cheap in terms of memory needed. Moreover, overfitting is less of a problem, resulting in better generalization.

These two observations have further implications: first, due to the reduced memory requirement, we can tackle original objects of space–bandwidth product (SBP) 128×128, higher than previously reported [28,29]. Second, the use of the convolutional architecture is counterintuitive because the scatterer may not be shift-invariant. Indeed, in Fig. 2 we show that it is not. It may seem justified, therefore, to worry whether the reduced memory and antioverfitting benefits of CNN may be outweighed. However, we found that the inverse estimate obtained by the CNN does in fact learn to compensate for the scatterer's shift variance, as shown in Fig. 12.

To characterize IDiffNet response, we conducted training and testing with well-calibrated diffusers of known grit size and well-calibrated intensity objects produced by a spatial light modulator. We also examined a large set of databases, including classes of objects with naturally embedded sparsity (e.g., handwritten characters or digits). These experiments enabled us to precisely quantify when IDiffNet requires strong sparsity constraints to become effective, as a function of diffuser severity. (the smaller the grit size, the more ill-posed the inverse problem becomes.)

The adoption of the NPCC instead of the more commonly used mean absolute error (MAE) as the loss function for training IDiffNet was an additional enabling factor in obtaining high-SBP image reconstructions through strong scatter. We compared the performance of these two loss functions under different imaging conditions and to different training databases determining the object priors that the networks learn, and showed that the NPCC is preferable for cases of relatively sparse objects (e.g., characters) and strong scatter. Lastly, we probed the interior of our trained IDiffNets through the well-established test of maximally activated patterns (MAPs) [37] and compared to standard denoising networks to eliminate the possibility that IDiffNet might be acting trivially instead of having learned anything about the diffuser and the objects' priors.

The structure of the paper is as follows: In Section 2, we describe the architecture of our computational imaging system, including the hardware and the IDiffNet machine learning architecture. In Section 3, the reconstruction results are analyzed, including the effects of scattering strength, object complexity (i.e., the object priors that the neural networks must learn), and choice of the loss function for training. In Section 4, we test the spatial resolution as well as the degree of shift invariance for the IDiffNets trained in different conditions. The comparison to a denoising neural network is described in Section 5, and concluding thoughts are in Section 6.

2. COMPUTATIONAL IMAGING SYSTEM ARCHITECTURE

The optical configuration that we consider in this paper is shown in Fig. 1. Light from a He–Ne laser source (Thorlabs, HNL210 L, 632.8 nm) is transmitted through a spatial filter, which consists of a microscope objective (Newport, M-60 ×, 0.85 NA) and a pinhole aperture ($D = 5$ μm). After being collimated by the lens ($f = 150$ mm), the light is reflected by a mirror and then passes through a linear polarizer, followed by

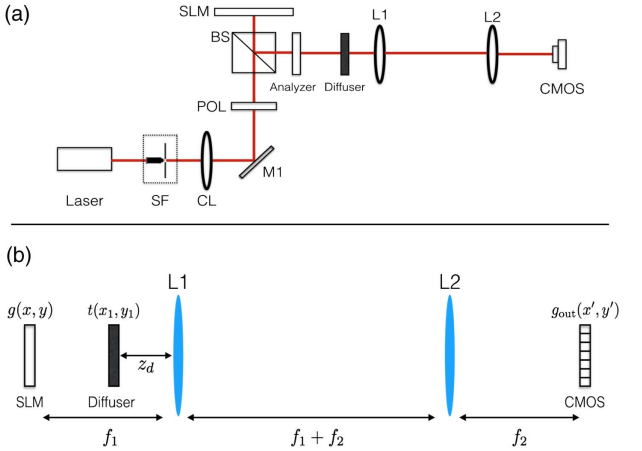

Fig. 1. Optical configuration. (a) Experimental arrangement. SF, spatial filter; CL, collimating lens; M, mirror; POL, linear polarizer; BS, beam splitter; SLM, spatial light modulator. (b) Detail of the telescopic imaging system.

a beam splitter. A spatial light modulator (Holoeye, LC-R 720, reflective) is placed normally incident to the transmitted light and acts as a pixel-wise intensity object. The SLM pixel size is 20 μm × 20 μm and number of pixels is 1280 × 768, out of which the central 512 × 512 portion only is used in the experiments. The SLM-modulated light is then reflected by the beam splitter and passes through a linear polarization analyzer before being scattered by a glass diffuser. A telescopic imaging system is built after the glass diffuser to image the SLM onto a complementary metal-oxide semiconductor (CMOS) camera (Basler, A504k), which has a pixel size of 12 μm × 12 μm. To match the pixel size of the CMOS with that of the SLM, we built the telescope using two lenses, L_1 and L_2, of focal lengths: $f_1 = 250$ mm and $f_2 = 150$ mm. As a result, the telescope magnifies the object by a factor of 0.6, which is consistent with the ratio between the pixel sizes of the CMOS and SLM. The total number of pixels on the CMOS is 1280 × 1024, but we crop only the central 512 × 512 square for processing; thus, the number of pixels measured by the CMOS camera, as well as their size, match 1:1 the object pixels at the SLM. Images recorded by the CMOS camera are then processed on an Intel i7 CPU. The neural network computations are performed on a GTX1080 graphics card (NVIDIA).

The modulation performance of the SLM depends on the orientations of the polarizer and analyzer. Here, we implement the cross-polarization arrangement to achieve a high-intensity modulation contrast. Specifically, we set the incident beam to be linearly polarized along the horizontal direction and also set the linear polarization analyzer to be oriented along the vertical direction. We experimentally calibrate the correspondence between the 8-bit grayscale input images projected onto the SLM and intensity modulation values of SLM (see Supplement 1, Section 1). We find that in this arrangement, the intensity modulation of the SLM follows a monotonic relationship with respect to assigned pixel value, and a maximum intensity modulation ratio of ∼17 can be achieved. At the same time, the SLM also introduces phase modulation, which is correlated with the intensity modulation due to the optical anisotropy of the liquid crystal molecules. The phase depth is $\sim 0.6\pi$. Fortunately, the influence of this phase modulation is negligible in the formation of the speckle images that we captured in this system; a detailed demonstration can be found in Section 2 of Supplement 1. Therefore, we are justified in treating this SLM as a pure-intensity object.

As shown in Fig. 1(b), the glass diffuser is inserted at a distance z_d in front of the lens L1. Here, we approximate the glass diffuser as a thin mask whose amplitude transmittance is $t(x_1, y_1)$. In this case, a forward model can be derived to relate the optical field at the detector plane $g_{\text{out}}(x', y')$ to the optical field at the object plane $g(x, y)$ (constant terms have been neglected) [38],

$$g_{\text{out}}(x',y') = \left\{ e^{\frac{-i\pi f_1^2}{\lambda(f_1-z_d)f_2^2}(x'^2+y'^2)} \cdot \int\int \mathrm{d}x\mathrm{d}y \left[g(x,y) e^{\frac{i\pi}{\lambda(f_1-z_d)}(x^2+y^2)} \cdot T\left(\frac{x+f_1x'/f_2}{\lambda(f_1-z_d)}, \frac{y+f_1y'/f_2}{\lambda(f_1-z_d)} \right) \right] \right\} * \left[\frac{J_1\left(\frac{2\pi R}{\lambda f_2}\sqrt{x'^2+y'^2}\right)}{\sqrt{x'^2+y'^2}} \right], \tag{2}$$

where λ is the light wavelength, R the radius of the lens L_2, and $J_1(\cdot)$ denotes the first-order Bessel function of the first kind. T is the Fourier spectrum of the diffuser, as follows: $T(u,v) = \iint \mathrm{d}x_1\mathrm{d}y_1[t(x_1,y_1)e^{-i2\pi(x_1u+y_1v)}]$. Here, * denotes the convolution product, and the last term in the convolution accounts for the influence of the finite aperture size of the lenses.

We model the diffuser transmittance $t(x_1, y_1)$ as a pure-phase random mask, i.e., $t(x_1,y_1) = \exp[\frac{i2\pi\Delta n}{\lambda} D(x_1,y_1)]$, where $D(x_1, y_1)$ is the random height of the diffuser surface, and Δn is the difference between the refractive indices of the diffuser and the surrounding air ($\Delta n \approx 0.52$ for glass diffusers). The random surface height $D(x_1, y_1)$ can be modeled as follows [39]:

$$D(x,y) = W(x,y) * K(\sigma). \tag{3}$$

Here, $W(x, y)$ is a set of random height values chosen according to the normal distribution at each discrete sample location (x, y), i.e., $W \sim N(\mu, \sigma_0)$; and $K(\sigma)$ is a zero-mean Gaussian smoothing kernel having full width half-maximum (FWHM) value of σ.

The values of μ, σ_0, and σ are determined by the grit size of the chosen glass diffuser [40]. In this paper, we use two glass diffusers of different grit size: 600-grit (Thorlabs, DG10-600-MD) and 220-grit (Edmund, 45-653). We have also conducted analysis and experiments with a 400-grit diffuser (OptoSigma, DFB1-30C02-400), included as part of the Supplement 1, and found the results almost identical to 220-grit size. Using the 220-grit and 600-grit values in Eqs. (2) and (3), we simulate the point spread function (PSF) of our imaging system as shown in Fig. 2, with a point source at the center of the object plane as input. We can see that the PSF for the 600-grit diffuser has a sharp peak at the center, while the PSF for the 220-grit diffuser spreads more widely. This indicates that the 220-grit diffuser scatters the light much more strongly than the 600-grit diffuser.

It is important to emphasize that, due to the existence of the diffuser, the imaging system is no longer shift-invariant. As can be seen in Eq. (2), the optical field at the detector plane g_{out} cannot be expressed as a convolution of the object g and a shift-invariant PSF term. The degree of shift variance may be compared using the PSF correlation function

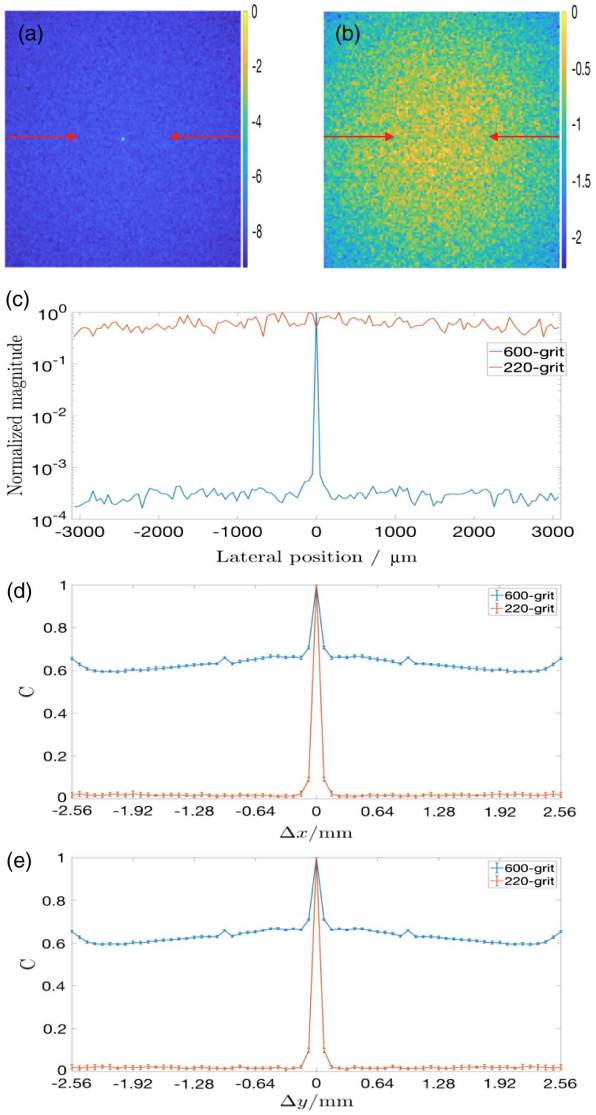

Fig. 2. Point spread functions (PSFs) and degree of shift variance of the imaging system. (a) PSF for the 600-grit diffuser: $\mu = 16$ μm, $\sigma_0 = 5$ μm, $\sigma = 4$ μm. (b) PSF for the 220-grit diffuser: $\mu = 63$ μm, $\sigma_0 = 14$ μm, $\sigma = 15.75$ μm. (c) Comparison of the profiles of the two PSFs along the lines indicated by the red arrows in (a) and (b). (d) Degree of shift variance along the x direction ($\Delta y = 0$). (e) Degree of shift variance along the y direction ($\Delta x = 0$). Other simulation parameters are set to be the same as the actual experiment: $z_d = 15$ mm, $R = 12.7$ mm, and $\lambda = 632.8$ nm. All the PSF plots are in logarithmic scale.

$$C(\Delta x, \Delta y) = \iiiint \langle h(x', y'; x, y) h(x', y'; x + \Delta x, y + \Delta y) \rangle \mathrm{d}x\mathrm{d}y\mathrm{d}x'\mathrm{d}y'. \tag{4}$$

Here, $h(x', y'; x, y)$ denotes the PSF on the detector plane (x', y') due to a point source in the object plane at location (x, y). Δx and Δy are the shifts in the object plane along x and y direction, respectively, and $\langle \cdot \rangle$ denotes the ensemble average over many simulated realizations of the diffuser. To make the comparison between different values of Δx, Δy possible, we normalized $h(.,.)$ to have zero mean and standard deviation equal to one.

Two slices of the PSF correlation function along the Δx and Δy directions, each for 10 random realizations of the simulated diffuser, are shown in Fig. 2(d) and Fig. 2(e), respectively, for the two grit sizes. As expected, in the 600-grit case, where scattering is weak, the shifted PSFs are more correlated than those in the 220-grit case. In both cases, the degree of correlation between the shifted PSFs decreases as the shift becomes larger. In addition, the degree of shift variance along the x direction is almost identical to that along the y direction.

We may also represent Eq. (2) in terms of a forward operator H_g: $g_{\text{out}}(x', y') = H_g g(x, y)$. When the object is pure-intensity, i.e., $g(x, y) = \sqrt{I(x, y)}$, the relationship between the raw intensity captured at the detector plane $I_{\text{out}}(x', y')$ and the object intensity $I(x, y)$ can also be represented in terms of another forward operator H: $I_{\text{out}}(x', y') = HI(x, y) = [SH_g Sr]I(x, y)$. Here, S denotes the modulus square operator, and Sr denotes the square root operator. Then, to reconstruct the intensity distribution of the object, we have to formulate an inverse operator H^{inv} such that

$$\hat{I}(x, y) = H^{\text{inv}} I_{\text{out}}(x', y'), \tag{5}$$

where $\hat{I}(x, y)$ is an acceptable estimate of the intensity object.

Owing to the randomness of H, it is difficult to obtain its explicit form and do the inversion accordingly; prior works referenced in Section 1 employed measurements of the scattering matrix to obtain H approximately. Here, we instead use IDiffNet, a deep neural network (DNN) trained to the underlying inverse mapping given a set of training data. IDiffNet uses the densely connected convolutional network (DenseNet) architecture [41], where each layer connects to every other layer within the same block in a feed-forward fashion. Compared to conventional convolutional networks, DenseNets have more direct connections between the layers, strengthening feature propagation, encouraging feature reuse, and substantially reducing the number of parameters. Therefore, DenseNets have better generalization capability.

A diagram of IDiffNet is shown in Fig. 3. The input to IDiffNet is the speckle pattern captured by the CMOS. It first passes through a dilated convolutional layer with filter size 5×5 and dilation rate 2, and is then successively decimated by six dense and downsampling transition blocks. After transmitting through another dense block, it successively passes through six dense and upsampling transition blocks and an additional upsampling transition layer. Finally, the signals pass through a standard convolutional layer with filter size 1×1, and the estimate of the object is produced. This is the "encoder–decoder network" architecture [42,43], where the dense and downsampling transition blocks serve as encoder to extract the feature maps from the input patterns, and the dense and upsampling transition blocks serve as decoder to perform pixel-wise regression. Owing to the scattering by the glass diffusers, the intensity at one pixel of the image plane is influenced by several nearby pixels at the object plane. Therefore, we use dilated convolutions with dilation rate 2 and a larger filter size of 5×5, compared to filter size 3×3 in [30], in all our dense blocks to increase the receptive field of the convolution filters. In addition, we also use skip connections [44] to pass high-frequency information learned in the initial layers down the network toward the output reconstruction. Additional details about the architecture and training of IDiffNet are provided in Section 3 of Supplement 1.

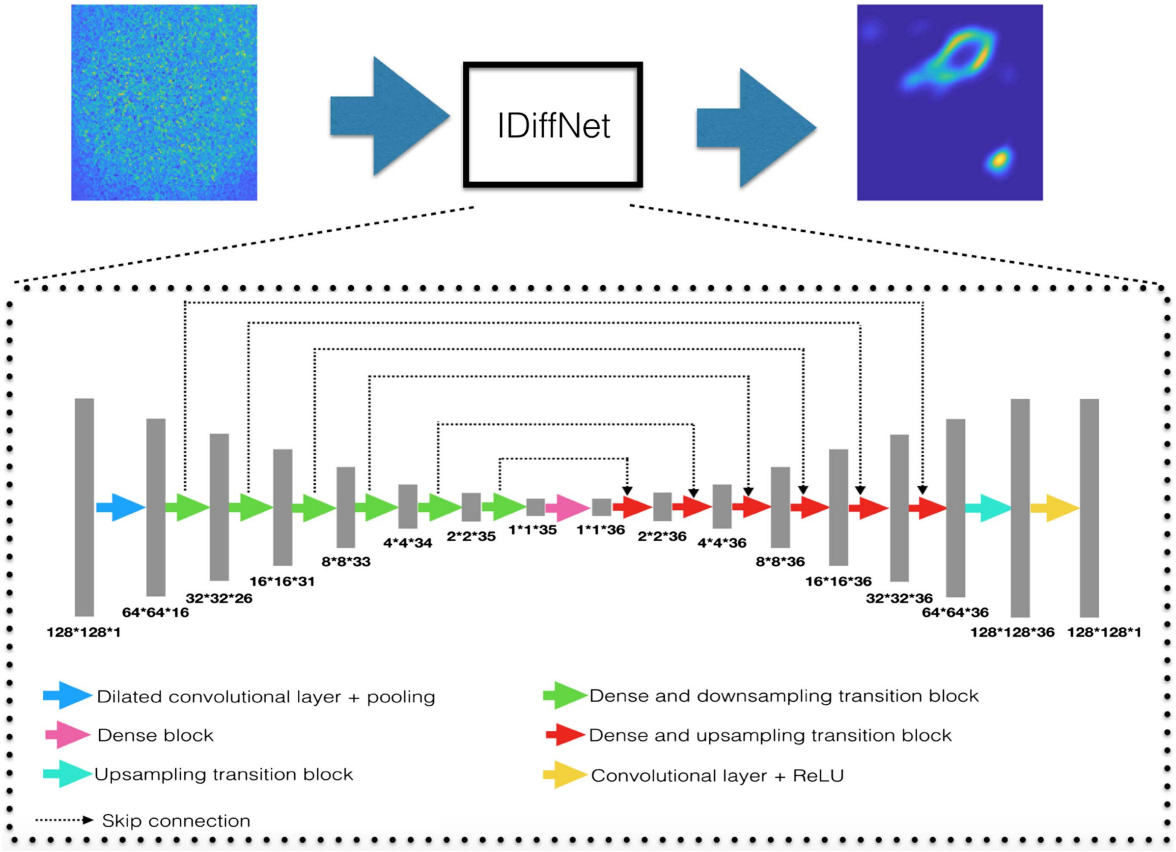

Fig. 3. IDiffNet, our densely connected neural network that images through diffuse media.

3. RESULTS AND NETWORK ANALYSIS

Our experiment consists of two phases: training and testing. During the training process, we randomly choose image samples from a training database. The space bandwidth products of the original images are all 128 × 128, and we magnify each image by a factor of 4 before uploading to the SLM. The corresponding speckle patterns are captured by the CMOS. As mentioned in Section 2, we crop only the central 512 × 512 square of the CMOS. We further downsample the captured speckle patterns by a factor of 4 and subtract from them a reference speckle pattern, which is obtained by uploading to the SLM a uniform image with all pixels equal to zero. The purpose of this subtraction operation is to eliminate the background noise on the CMOS and also to better extract differences between speckle patterns resulting from different objects.

After the subtraction operation, we feed the resulting speckle patterns into IDiffNet for training. In this way, the input and output signal dimensions are both 128 × 128. We collected data from six separate experiment runs: each time we used training inputs from one of the three different databases—Faces-LFW [45], ImageNet [46], or MNIST [47]—and inserted one of the two glass diffusers that we have into the imaging system. Each of our training dataset consists of 10,000 object-speckle pattern pairs. These data were used to train six separate IDiffNets for evaluation. In the testing process, we sample disjoint examples from the same database (Faces-LFW, ImageNet, or MNIST) and other databases such as Characters, CIFAR [48], and Faces-ATT [49]. Altogether, 450 examples are used in the test dataset, including 50 Characters, 40 Faces-ATT, 60 CIFAR, 100 MNIST, 100 Faces-LFW, and 100 ImageNet. We upload these test examples to the SLM and capture their corresponding speckle patterns using the same glass diffuser as the training phase. We then input these speckle patterns to our trained IDiffNet and compare the output to the ground truth.

In training the IDiffNets, we use two different loss functions and compare their performances. The first loss function that we consider is the MAE, defined as follows:

$$\mathrm{MAE} = \frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h}|Y(i,j) - G(i,j)|. \tag{6}$$

Here, w, h are the width and height of the output, Y is the output of the last layer, and G is the ground truth.

The qualitative and quantitative reconstruction results when using MAE as the loss function are shown in Fig. 4 and Fig. 5, respectively. From Fig. 4, we find that, generally, IDiffNet's reconstruction performance for the 600-grit diffuser is better than that for the 220-grit diffuser. High-quality reconstructions are achieved for the 600-grit diffuser when IDiffNets are trained on Faces-LFW (column iv) and ImageNet (column v). For the 220-grit diffuser, the best reconstruction is obtained when IDiffNet is trained on the ImageNet database (column ix). The recovered images are close to the low-pass filtered version of the original image, where we can visualize the general shape (salient features) but the high frequency features are missing. This result is expected since the scattering caused by the 220-grit diffuser is much stronger than that of the 600-grit diffuser, as we had already

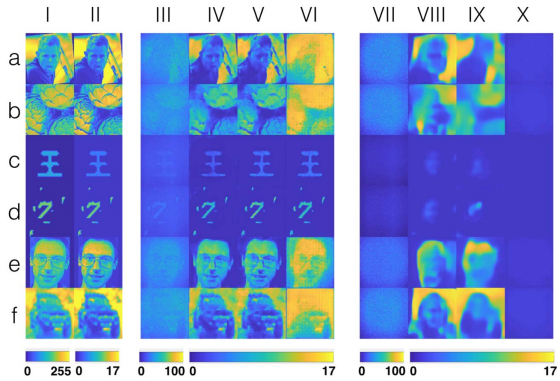

Fig. 4. Qualitative analysis of IDiffNet trained using MAE as the loss function. (i) Ground truth pixel value inputs to the SLM. (ii) Corresponding intensity images calibrated by SLM response curve. (iii) Raw intensity images captured by CMOS detector for 600-grit glass diffuser. (iv) IDiffNet reconstruction from raw images when trained using Faces-LFW dataset [45]. (v) IDiffNet reconstruction when trained using ImageNet dataset [46]. (vi) IDiffNet reconstruction when trained using MNIST dataset [47]. Columns (vii)–(x) follow the same sequence as (iii)–(vi), but in these sets the diffuser used is 220-grit. Rows (a)–(f) correspond to the dataset from which the test image is drawn: (a) Faces-LFW, (b) ImageNet, (c) Characters, (d) MNIST, (e) Faces-ATT [49], (f) CIFAR [48], respectively.

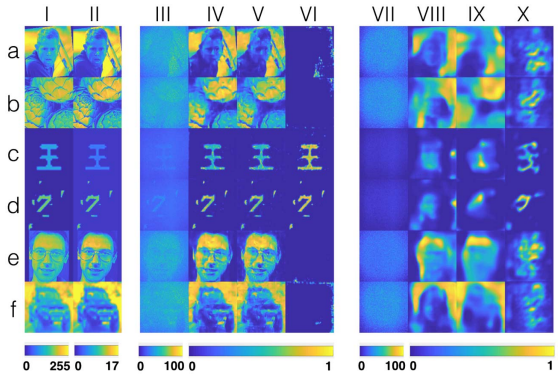

Fig. 6. Qualitative analysis of IDiffNets trained using NPCC as the loss function. (i) Ground truth pixel value inputs to the SLM. (ii) Corresponding intensity images calibrated by SLM response curve. (iii) Raw intensity images captured by CMOS detector for 600-grit glass diffuser. (iv) IDiffNet reconstruction from raw images when trained using Faces-LFW dataset [45]. (v) IDiffNet reconstruction when trained using ImageNet dataset [46]. (vi) IDiffNet reconstruction when trained using MNIST dataset [47]. Columns (vii)–(x) follow the same sequence as (iii)–(vi) but in these sets the diffuser used is 220-grit. Rows (a)–(f) correspond to the dataset from which the test image is drawn: (a) Faces-LFW, (b) ImageNet, (c) Characters, (d) MNIST, (e) Faces-ATT [49], (f) CIFAR [48], respectively.

deduced from Fig. 2. As a result, we can still visualize some features of the object in the raw intensity image captured in the 600-grit diffuser case. By contrast, what we capture in the 220-grit diffuser case looks indistinguishable from pure speckle, without any object details visible. This means we should expect it to be much more difficult for IDiffNet to do the inversion.

Noticeable from Fig. 4 is that when IDiffNet is trained on MNIST for the 220-grit diffuser (column x), all the reconstructions seem to be uniform. The reason is that the objects that this IDiffNet was trained on were sparse; hence, it also tends to make sparse estimates. Unfortunately, in this case the sparse local minima where IDiffNet is trapped are featureless. Tackling this problem motivated us to examine the NPCC as an alternative loss function.

The NPCC is defined as follows [50]:

$$\mathrm{NPCC}=\frac{-1\times\sum_{i=1}^{w}\sum_{j=1}^{h}(Y(i,j)-\tilde{Y})(G(i,j)-\tilde{G})}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(Y(i,j)-\tilde{Y})^2}\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(G(i,j)-\tilde{G})^2}}. \tag{7}$$

Here, $\tilde{G}$ and $\tilde{Y}$ are the mean value of the ground truth G and the IDiffNet output Y, respectively.

The qualitative and quantitative reconstruction results using the NPCC as the loss function are shown in Fig. 6 and Fig. 7, respectively. The reconstructed images are normalized since the NPCC value will be the same if we multiply the reconstruction by any positive constants. Similar to the case where MAE is used as the loss function, the reconstruction is better in the 600-grit

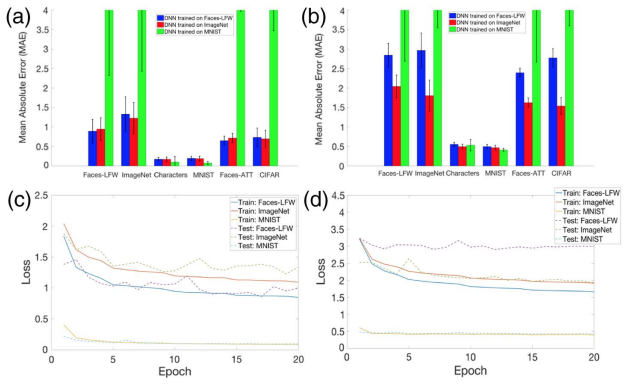

Fig. 5. Quantitative analysis of IDiffNet trained using MAE as the loss function. Test errors for IDiffNet trained on Faces-LFW (blue), ImageNet (red), and MNIST (green) on six datasets when the diffuser used is (a) 600-grit and (b) 220-grit. The training and testing error curves when the diffuser used is (c) 600-grit and (d) 220-grit.

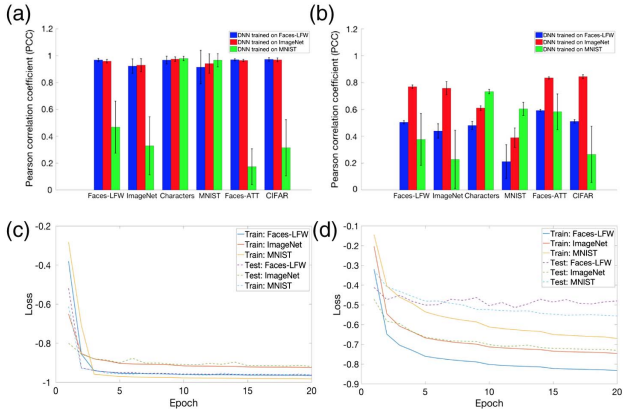

Fig. 7. Quantitative analysis of our trained deep neural networks for using NPCC as the loss function. Test errors for the IDiffNets trained on Faces-LFW (blue), ImageNet (red), and MNIST (green) on six datasets when the diffuser used is (a) 600-grit and (b) 220-grit. The training and testing error curves when the diffuser used is (c) 600-grit and (d) 220-grit.

diffuser case than the 220-grit diffuser case. However, when IDiffNet is trained on MNIST for the 220-grit diffuser (column x), high-quality reconstruction is achieved for the test images coming from Characters and MNIST database (row c and d). This is in contrast to the MAE-trained case, thus indicating that NPCC is a more appropriate loss function to use in this case. It helps IDiffNet to learn the sparsity in the ground truth and in turn use the sparsity as a strong prior for estimating the inverse. In addition, when trained on ImageNet for the 220-grit diffuser (column ix), IDiffNet is still able to reconstruct the general shape (salient features) of the object. But the NPCC-trained reconstructions are visually slightly worse compared to the MAE-trained cases.

In both MAE and NPCC training cases, IDiffNet performance also depends on the dataset that it is trained on. From Figs. 4 and 6, we observe that IDiffNet generalizes best when being trained on ImageNet and has the most severe overfitting problem when being trained on MNIST. Specifically, when IDiffNet is trained on MNIST, even for the 600-grit diffuser (column vi), it works well if the test image comes from the same database or a database that shares the same sparse characteristics as MNIST (e.g., characters). It gives much worse reconstruction when the test image comes from a much different database. When IDiffNet is trained on Faces-LFW, it generalizes well for the 600-grit diffuser, but for the 220-grit diffuser it exhibits overfitting: it tends to reconstruct a face at the central region, as in Horisaki's case. When IDiffNet is trained on ImageNet, it generalizes well even for the 220-grit diffuser. As we can see in column ix, for all the test images, IDiffNet is able to at least reconstruct the general shapes (salient features) of the objects. This indicates that IDiffNet has learned at the very least a generalizable mapping of low-level textures between the captured speckle patterns and the input images. Similar observation may also be made from Figs. 5 and 7. From subplots (a) and (b) in both figures, we notice that the IDiffNets trained on MNIST have much higher MAEs/lower PCCs when tested on other databases. As shown in subplot (d), the IDiffNets trained on Faces-LFW have a large discrepancy between training and test error, while for IDiffNets trained on ImageNet, the training and testing curves converge to almost the same level. An explanation for this phenomenon is that all the images in MNIST or Faces-LFW databases share the same characteristics (e.g., sparse, circular shape), imposing a strong prior on IDiffNet. On the other hand, the ImageNet database consists of a mixture of generic images that have not too much in common. As a result, IDiffNet trained on ImageNet generalizes better. It is worth noting that overfitting in our case evidences itself as face-looking "ghosts" occurring when IDiffNet trained on Faces-LFW tries to reconstruct other kinds of images, for example (see Fig. 6, column viii). This is again similar to Horisaki's observations [28].

From comparing the four possible combinations of weak to strong scattering and constrained dataset (e.g., MNIST) to generic dataset (e.g., ImageNet), we conclude the following: when scattering is weak, it is to our benefit to train the IDiffNets on a generic dataset because the resulting neural networks generalize better and can cope with the scattering also for general test images. On the other hand, when scattering is strong, it is beneficial to use a relatively constrained dataset with strong sparsity present in the typical objects: the resulting neural networks are then more prone to overfitting, but now this works to our benefit for overcoming strong scattering (at the cost, of course, of working only for test objects coming from the more restricted database). The choice of optimization functional makes this tradeoff even starker: MAE apparently does not succeed in learning the strong sparsity even for MNIST datasets, whereas the NPCC does much better, even being capable of reconstructing test objects under the most severe scattering conditions (220-grit diffuser, column x in Fig. 6) as long as the objects are drawn from the sparse dataset. These observations are summarized in Table 1.

4. RESOLUTION AND SHIFT INVARIANCE TESTS FOR IDIFFNET

In this section, we investigate the spatial resolution of our trained IDiffNet. Without the diffuser, our system is a telescope of

Table 1. Summary of Reconstruction Results in Different Cases

		600-grit		220-grit	
	Training Dataset	**Loss: MAE**	**Loss: NPCC**	**Loss: MAE**	**Loss: NPCC**
Test: Faces-LFW	Faces-LFW	✓	✓	•	×
	ImageNet	✓	✓	•	•
	MNIST	×	×	×	×
Test: ImageNet	Faces-LFW	✓	✓	×	×
	ImageNet	✓	✓	•	•
	MNIST	×	×	×	×
Test: Characters	Faces-LFW	✓	✓	×	×
	ImageNet	✓	✓	•	•
	MNIST	✓	✓	×	✓
Test: MNIST	Faces-LFW	✓	✓	×	×
	ImageNet	✓	✓	•	•
	MNIST	✓	✓	×	✓
Test: Faces-ATT	Faces-LFW	✓	✓	×	×
	ImageNet	✓	✓	•	•
	MNIST	×	×	×	×
Test: CIFAR	Faces-LFW	✓	✓	×	×
	ImageNet	✓	✓	•	•
	MNIST	×	×	×	×

[✓: visually recognizable; •: salient feature recognizable; ×: visually unrecognizable.]

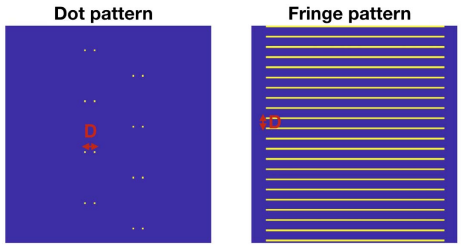

Fig. 8. Resolution test patterns. Left: dot pattern. Right: fringe pattern.

numerical aperture NA $= 12.7/250 = 0.0508$. The diffraction-limited Abbé resolution is $d_0 = \lambda/(2\text{NA}) = 6.23$ μm.

With the 600-grit diffuser in the system, we analyze experimentally four IDiffNets that we trained on either the ImageNet or MNIST database and using either MAE or the NPCC as the loss function. To evaluate the spatial resolution, we designed two different sets of test patterns, dots and fringes, as shown in Fig. 8. The dots are constructed as superpixels from 4 × 4 pixels (80 μm × 80 μm), and the fringes are constructed as bands of width equal to 4 pixels (80 μm). These choices make the dot and fringe spacings consistent with the sampling scheme chosen in the experiments of Section 3. It should also be mentioned that the superpixel size places a limit on IDiffNet resolution, since IDiffNet is trained with examples whose sampling distance is equal to one superpixel. In the dot pattern set, each pattern contains eight dot pairs, and the spacings D between the two dots within the same pair are set to be the same. The entire set consists of 10 such dot patterns with D gradually varying from 1 superpixel to 10 superpixels. In the fringe pattern set, each pattern contains equally spaced fringes. Similarly, the entire set consists of 10 such fringe patterns with the spacing D gradually varying from 1 superpixel to 10 superpixels.

Those resolution test patterns are displayed on the SLM, and the corresponding speckle patterns are captured and fed into our trained IDiffNet for reconstruction. The results are shown in Figs. 9 and 10. Here, we show only the resolution test results of the IDiffNets trained using MAE as the loss function. The choice of the loss function actually does not affect the spatial resolution of the trained IDiffNet in the 600-grit diffuser case. The resolution test results of the IDiffNets trained using NPCC as the loss function can be found in Section 4 of Supplement 1.

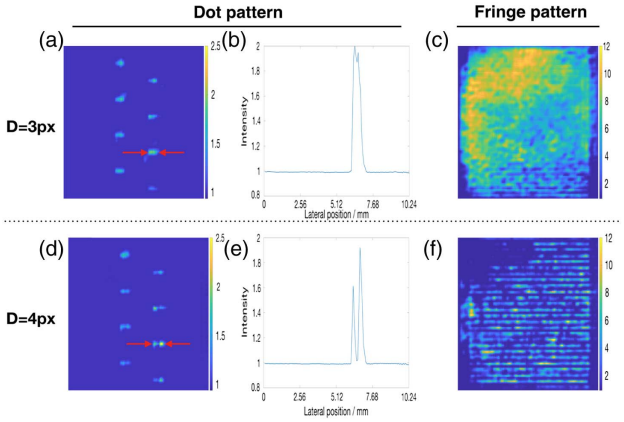

Fig. 9. Experimental resolution test result for IDiffNet trained on MNIST using MAE as loss function. The diffuser used is 600-grit. (a) Reconstructed dot pattern when $D = 3$ superpixels. (b) 1D cross-section plot along the line indicated by red arrows in (a). (c) Reconstructed fringe pattern when $D = 3$ superpixels. (d) Reconstructed dot pattern when $D = 4$ superpixels. (e) 1D cross-section plot along the line indicated by red arrows in (d). (f) Reconstructed fringe pattern when $D = 4$ superpixels.

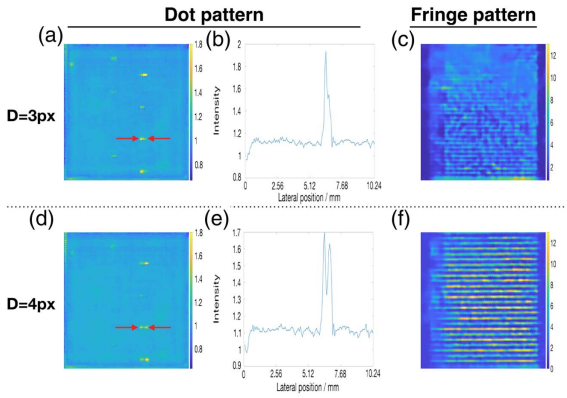

Fig. 10. Experimental resolution test result for IDiffNet trained on ImageNet using MAE as loss function. The diffuser used is 600-grit. (a) Reconstructed dot pattern when $D = 3$ superpixels. (b) 1D cross-section plot along the line indicated by red arrows in (a). (c) Reconstructed fringe pattern when $D = 3$ superpixels. (d) Reconstructed dot pattern when $D = 4$ superpixels. (e) 1D cross-section plot along the line indicated by red arrows in (d). (f) Reconstructed fringe pattern when $D = 4$ superpixels.

As shown in Fig. 9, the IDiffNet trained on the MNIST database is able to resolve two dots with spacing $D = 4$ superpixels but fails to distinguish two dots with spacing $D = 3$ superpixels. The same spatial resolution is demonstrated using fringe patterns as well, where nearby fringes with spacing $D = 4$ superpixels are resolved while fringes with spacing $D = 3$ superpixels are unable to be distinguished. In addition, we find that the reconstruction qualities of dot patterns are better than those of the fringe patterns. This result is as expected since the MNIST training database imposes a strong sparsity prior in a set of basis functions that themselves look relatively spatially sparse [51]. This property makes IDiffNet perform better on spatially sparse test samples (dot patterns) than other less sparse test samples (fringe patterns). Therefore, dot patterns are more appropriate to be used to test the resolution of IDiffNet trained on MNIST.

For the IDiffNet trained on ImageNet, its spatial resolution is the same as the MNIST training case, as demonstrated in Fig. 10. However, the reconstruction qualities of fringe patterns are better than those of the dot patterns since the ImageNet training database contains more general images, which are sparse in a set of basis functions that is spatially richer than the MNIST dictionary [52]. Because of this observation, fringe patterns are more appropriate for testing the resolution of IDiffNet trained on ImageNet.

Now, let us test the spatial resolution of IDiffNet trained using the 220-grit diffuser. In this strong scattering case, as described in Section 3, we have to use MNIST as the training database and use the NPCC as the loss function. To match the strong prior imposed by the MNIST database, we design the resolution test pattern as shown in Fig. 11(a), where those dots are placed in a layout resembling the digit "7" and the spacing between nearby dots is

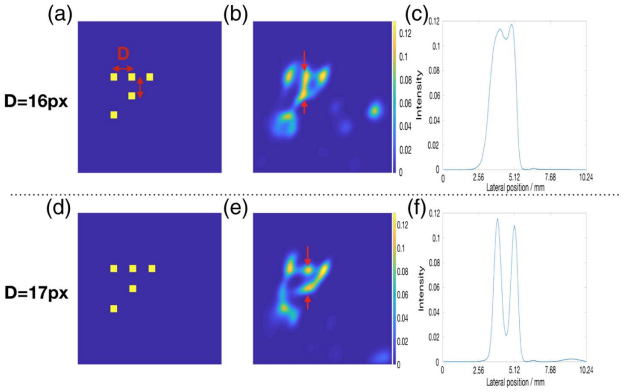

Fig. 11. Experimental resolution test result for IDiffNet trained on MNIST using NPCC as loss function. The diffuser used is 220-grit. (a) Resolution test pattern when $D = 16$ superpixels. (b) Reconstructed test pattern when $D = 16$ superpixels. (c) 1D cross-section plot along the line indicated by red arrows in (b). (d) Resolution test pattern when $D = 17$ superpixels. (e) Reconstructed test pattern when $D = 17$ superpixels. (f) 1D cross-section plot along the line indicated by red arrows in (e).

D. The entire set consists of 10 such patterns with the spacing D gradually varying from 10 superpixels to 19 superpixels. As shown in Fig. 11, we can find that the trained IDiffNet is able to resolve nearby dots with spacing $D = 17$ superpixels but fails to distinguish two dots with spacing $D = 16$ superpixels. As expected, the spatial resolution in this case is worse than that in the 600-grit diffuser case.

In light of our earlier observation about limited shift invariance in the forward operator (see discussion after Fig. 2), we also analyzed IDiffNet's shift invariance. In simulation, as in Section 2, we chose 100 test images in the MNIST database. For each image $I(x, y)$, we first obtain its corresponding speckle pattern $I_{\text{out}}(x', y'; x, y)$. Then, we shift $I(x, y)$ along the x direction for some distance Δx and obtain the corresponding speckle pattern $I_{\text{out}}(x', y'; x + \Delta x, y)$. After that, we shift the speckle pattern $I_{\text{out}}(x', y'; x + \Delta x, y)$ back by a distance $\Delta x' = m\Delta x$ to obtain $I_{\text{out}}(x' - \Delta x', y'; x + \Delta x, y)$, where $m = 0.6$ is the ratio between the camera and SLM pixel sizes. Finally, we compute the correlations in the speckle patterns as

$$C_s(\Delta x) = \text{PCC}[I_{\text{out}}(x', y'; x, y), I_{\text{out}}(x' - \Delta x', y'; x + \Delta x, y)]. \tag{8}$$

Here, PCC is defined as in Eq. (7) and without the negative sign.

Similarly, we can compute the correlations in the reconstructions $\hat{I}(x, y; x, y)$ as

$$C_r(\Delta x) = \text{PCC}[\hat{I}(x, y; x, y), \hat{I}(x - \Delta x, y; x + \Delta x, y)]. \tag{9}$$

As shown in Fig. 12, shift invariance after IDiffNet increases ($C_r > C_s$), demonstrating that IDiffNet has learned to compensate for shift variance in the forward operator. In fact, the domain of shift invariance obtained with IDiffNet is bigger than generally obtained by approaches based on the memory effect [13,16]. In the latter, the field of view (FOV) is limited by shift invariance in the forward operator, i.e., the domain of high correlation between shifted PSFs, and is typically small, e.g., FOV $\sim 2.5 \times 10^{-2}$ in [13]. On the other hand, our experimental results with

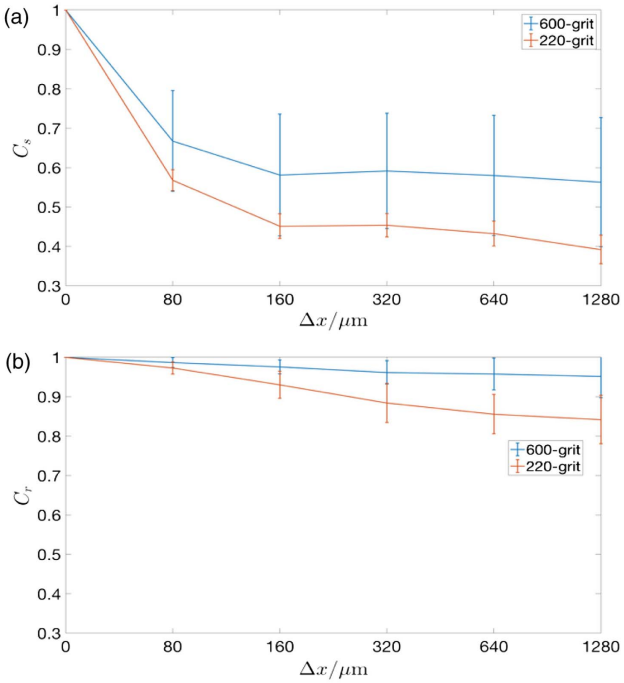

Fig. 12. Simulated shift invariance test. (a) Correlations in the speckle patterns C_s calculated on MNIST database. (b) Correlations in the reconstructions C_r calculated on MNIST database. In the 600-grit case, the IDiffNet is trained on ImageNet using MAE loss function; in the 220-grit case, the IDiffNet is trained on MNIST using NPCC loss function.

IDiffNet demonstrate FOV $\gtrsim 4 \times 10^{-2}$ for the same diffuser of 220-grit size.

5. COMPARISON WITH DENOISING NEURAL NETWORKS

To get a sense of what IDiffNets learn, we first compare their reconstruction results to those of a denoising neural network. Specifically, we use ImageNet as our training database. To each image in the training dataset, we simulate a noisy image using Poisson noise and make the peak signal-to-noise ratio (PSNR) of the resulting noisy image visually comparable to that of the corresponding speckle image captured using the 600-grit diffuser. We use Poisson noise rather than different kinds of noise such as Gaussian because Poisson noise is signal-dependent, similar to the diffuser case. We then train a denoising neural network using those noisy images. For the denoising neural network, we implement the residual network architecture [53]. Finally, we feed the test speckle images captured using the 600-grit diffuser into this denoising neural network and compare the outputs to those reconstructed by IDiffNet (using MAE as the loss function).

The comparison results are shown in Fig. 13. From column iv, we find that the denoising neural network works well when the test images are indeed noisy according to the Poisson model. However, as shown in column v, if we input the diffuse image into the denoising network, it can only output a highly blurred image, much worse than IDiffNet given the same diffuse input, as can be seen in column vi. This result demonstrates that IDiffNet is not doing denoising, although the speckle image obtained using the 600-grit diffuser visually looks similar to a noisy image.

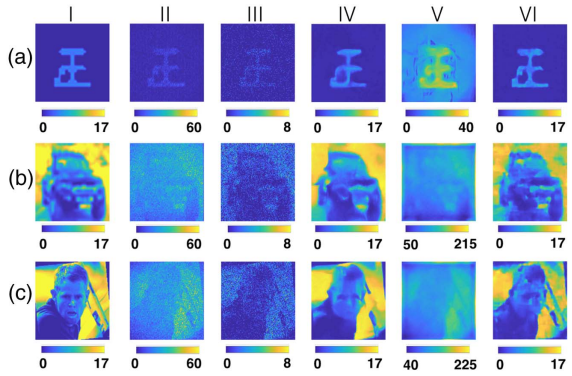

Fig. 13. Comparison between IDiffNets and a denoising neural network. (i) Ground truth intensity images calibrated by SLM response curve. (ii) Speckle images that we captured using the 600-grit diffuser (after subtracting the reference pattern). (iii) Noisy images generated by adding Poisson noise to the ground truth. (iv) Reconstructions of the denoising neural network when inputting the noisy image in (iii). (v) Reconstructions of the denoising neural network when inputting the speckle image in (ii). (vi) IDiffNet reconstructions when inputting the speckle image in (ii). (The images shown in column vi are the same as those in the column v of Fig. 4, duplicated here for the readers' convenience.) Rows (a)–(c) correspond to the dataset from which the test images are drawn: (a) Characters, (b) CIFAR [48], (c) Faces-LFW [45], respectively.

We posit the reason for this as follows: Poisson noise operates pixel-wise. Consequently, denoising for Poisson noise is effectively another pixel-wise operation that does not depend much on spatial neighborhood, except to the degree that applying priors originating from the structure of the object helps to denoise severely affected signals. A denoising neural network, then, learns spatial structure only as a prior on the class of objects it is trained on. However, this is not the case when imaging through a diffuser: then every pixel value in the speckle image is influenced by a set of nearby pixels in the original image. This may also be seen from the PSF plots shown in Fig. 2. The denoising neural network fails because it has not learned the spatial correlations between the nearby pixels and the correct kernel of our imaging system, as our IDiffNet has.

We also examined the MAPs of the IDiffNets and the denoising neural network; i.e., what types of inputs would maximize network filter response (gradient descent on the input with average filter response as loss function) [37]. Figure 14 shows the MAP analysis of two convolutional layers at different depths for all three neural networks. For both the shallow and deep layers, we find the MAPs of our IDiffNets are qualitatively different from those of the denoising network. This further corroborates that IDiffNet is not merely doing denoising. In addition, the MAPs of the 600-grit IDiffNet show finer textures compared to those of the 220-grit IDiffNet, indicating that the IDiffNet learns better in the 600-grit diffuser case.

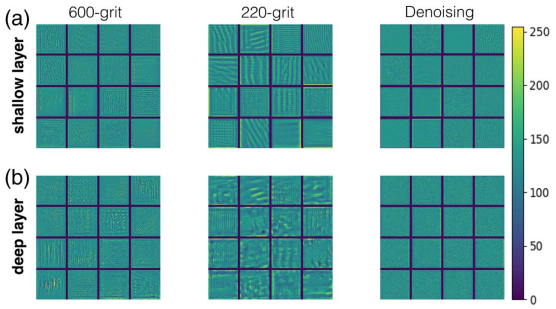

Fig. 14. Maximally activated patterns (MAPs) for different DNNs. (a) 128 × 128 inputs that maximally activate the filters in the convolutional layer at depth 5. (b) 128 × 128 inputs that maximally activate the filters in the convolutional layer at depth 13. (There are actually more than 16 filters at each convolutional layer, but we show only the 16 filters that have the highest activations here.)

6. CONCLUSIONS

We have demonstrated that IDiffNets, built according to the densely connected convolutional neural network architecture, can be used as an end-to-end approach for imaging through scattering media. The reconstruction performance depends on the scattering strength of the diffusers, the type of the training dataset (in particular, its sparsity), and the loss function used for optimization. The IDiffNets seem to learn automatically the properties of the scattering media, including the degree of shift invariance and how to at least partially compensate it, as well as the priors restricting the objects where the network is supposed to perform well, depending on the database the network was trained with.

Funding. Singapore-MIT Alliance for Research and Technology Centre (SMART); RAVEN Program, Intelligence Advanced Research Projects Activity (IARPA); U.S. Department of Energy (DOE) (DE-FG02-97ER25308).

Acknowledgment. Justin Lee acknowledges funding from the U.S. Department of Energy Computational Science Graduate Fellowship (CSGF). We are grateful to the anonymous reviewers for constructive suggestions and criticism.

See Supplement 1 for supporting content.

REFERENCES

1. V. I. Tatarski, *Wave Propagation in a Turbulent Medium* (Courier Dover, 2016).
2. A. Ishimaru, *Wave Propagation and Scattering in Random Media* (Academic, 1978), Vol. **2**.
3. W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," Science **248**, 73–76 (1990).
4. L. Moreaux, O. Sandre, and J. Mertz, "Membrane imaging by second-harmonic generation microscopy," J. Opt. Soc. Am. B **17**, 1685–1694 (2000).
5. S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy," Opt. Lett. **19**, 780–782 (1994).
6. M. G. Gustafsson, "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy," J. Microsc. **198**, 82–87 (2000).
7. T. Wilson, "Optical sectioning in fluorescence microscopy," J. Microsc. **242**, 111–116 (2011).
8. D. Lim, K. K. Chu, and J. Mertz, "Wide-field fluorescence sectioning with hybrid speckle and uniform-illumination microscopy," Opt. Lett. **33**, 1819–1821 (2008).
9. S. Popoff, G. Lerosey, R. Carminati, M. Fink, A. Boccara, and S. Gigan, "Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media," Phys. Rev. Lett. **104**, 100601 (2010).
10. S. Popoff, G. Lerosey, M. Fink, A. Boccara, and S. Gigan, "Image transmission through an opaque material," Nat. Commun. **1**, 81 (2010).

11. A. Drémeau, A. Liutkus, D. Martina, O. Katz, C. Schülke, F. Krzakala, S. Gigan, and L. Daudet, "Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques," Opt. Express **23**, 11898–11911 (2015).
12. J. Bertolotti, E. G. van Putten, C. Blum, A. Lagendijk, W. L. Vos, and A. P. Mosk, "Non-invasive imaging through opaque scattering layers," Nature **491**, 232–234 (2012).
13. O. Katz, P. Heidmann, M. Fink, and S. Gigan, "Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations," Nat. Photonics **8**, 784–790 (2014).
14. N. Stasio, C. Moser, and D. Psaltis, "Calibration-free imaging through a multicore fiber using speckle scanning microscopy," Opt. Lett. **41**, 3078–3081 (2016).
15. A. Porat, E. R. Andresen, H. Rigneault, D. Oron, S. Gigan, and O. Katz, "Widefield lensless imaging through a fiber bundle via speckle correlations," Opt. Express **24**, 16835–16855 (2016).
16. G. Osnabrugge, R. Horstmeyer, I. N. Papadopoulos, B. Judkewitz, and I. M. Vellekoop, "Generalized optical memory effect," Optica **4**, 886–892 (2017).
17. S. Feng, C. Kane, P. A. Lee, and A. D. Stone, "Correlations and fluctuations of coherent wave transmission through disordered media," Phys. Rev. Lett. **61**, 834–837 (1988).
18. I. Freund, M. Rosenbluh, and S. Feng, "Memory effects in propagation of optical waves through disordered media," Phys. Rev. Lett. **61**, 2328–2331 (1988).
19. E. Akkermans and G. Montambaux, *Mesoscopic Physics of Electrons and Photons* (Cambridge University, 2007).
20. R. W. Gerchberg, "A practical algorithm for the determination of the phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
21. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. **3**, 27–29 (1978).
22. Y. Park, W. Choi, Z. Yaqoob, R. Dasari, K. Badizadegan, and M. S. Feld, "Speckle-field digital holographic microscopy," Opt. Express **17**, 12285–12292 (2009).
23. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "DiffuserCam: lensless single-exposure 3D imaging," Optica **5**, 1–9 (2018).
24. U. Grenander, *General Pattern Theory—A Mathematical Study of Regular Structures* (Clarendon, 1993).
25. E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory **52**, 489–509 (2006).
26. D. J. Brady, K. Choi, D. L. Marks, R. Horisaki, and S. Lim, "Compressive holography," Opt. Express **17**, 13040–13049 (2009).
27. H.-Y. Liu, E. Jonas, L. Tian, J. Zhong, B. Recht, and L. Waller, "3D imaging in volumetric scattering media using phase-space measurements," Opt. Express **23**, 14461–14471 (2015).
28. R. Horisaki, R. Takagi, and J. Tanida, "Learning-based imaging through scattering media," Opt. Express **24**, 13738–13743 (2016).
29. M. Lyu, H. Wang, G. Li, and G. Situ, "Exploit imaging through opaque wall via deep learning," arXiv:1708.07881 (2017).
30. A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," Optica **4**, 1117–1125 (2017).
31. Y. Rivenson, Y. Zhang, H. Gunaydin, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," arXiv:1705.04286 (2017).
32. Y. Rivenson, Z. Gorocs, H. Gunaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," arXiv:1705.04709 (2017).
33. K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," IEEE Trans. Image Process. **26**, 4509–4522 (2017).
34. G. Satat, M. Tancik, O. Gupta, B. Heshmat, and R. Raskar, "Object classification through scattering media with deep learning on time resolved measurement," Opt. Express **25**, 17466–17479 (2017).
35. R. Horstmeyer, R. Y. Chen, B. Kappes, and B. Judkewitz, "Convolutional neural networks that teach microscopes how to image," arXiv:1709.07223 (2017).
36. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks* (1995), Vol. **3361**.
37. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Springer, 2014), pp. 818–833.
38. J. W. Goodman, *Introduction to Fourier Optics* (Roberts and Company, 2005).
39. N. Antipa, S. Necula, R. Ng, and L. Waller, "Single-shot diffuser-encoded light field imaging," in *IEEE International Conference on Computational Photography (ICCP)* (IEEE, 2016), pp. 1–11.
40. https://www.unc.edu/~rowlett/units/scales/grit.html.
41. G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," arXiv:1608.06993 (2016).
42. X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems* (2016), pp. 2802–2810.
43. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: a deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2481–2495 (2017).
44. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015), pp. 234–241.
45. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Technical Report 07-49 (University of Massachusetts, 2007).
46. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).
47. Y. LeCun, C. Cortes, and C. J. Burges, "MNIST handwritten digit database," AT&T Labs, 2010, http://yann.lecun.com/exdb/mnist.
48. A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical Report (University of Toronto, 2009).
49. F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of the Second IEEE Workshop on Applications of Computer Vision* (IEEE, 1994), pp. 138–142.
50. A. M. Neto, A. C. Victorino, I. Fantoni, D. E. Zampieri, J. V. Ferreira, and D. A. Lima, "Image processing using Pearson's correlation coefficient: applications on autonomous robotics," in *13th International Conference on Autonomous Robot Systems (Robotica)* (IEEE, 2013), pp. 1–6.
51. C. Gehring and S. Lemay, "Sparse coding," sibi **1**, 1 (2012).
52. B.-S. Kim, J. Y. Park, A. C. Gilbert, and S. Savarese, "Hierarchical classification of images by sparse approximation," Image Vision Comput. **31**, 982–991 (2013).
53. T. Remez, O. Litany, R. Giryes, and A. M. Bronstein, "Deep convolutional denoising of low-light images," arXiv:1701.01687 (2017).