Check for updates

OPEN

Image reconstruction through a multimode fiber with a simple neural network architecture

Changyan Zhu[1,4], Eng Aik Chan[2,4], You Wang[1], Weina Peng[3], Ruixiang Guo[2], Baile Zhang[1,2✉], Cesare Soci[1,2✉] & Yidong Chong[1,2✉]

Multimode fibers (MMFs) have the potential to carry complex images for endoscopy and related applications, but decoding the complex speckle patterns produced by mode-mixing and modal dispersion in MMFs is a serious challenge. Several groups have recently shown that convolutional neural networks (CNNs) can be trained to perform high-fidelity MMF image reconstruction. We find that a considerably simpler neural network architecture, the single hidden layer dense neural network, performs at least as well as previously-used CNNs in terms of image reconstruction fidelity, and is superior in terms of training time and computing resources required. The trained networks can accurately reconstruct MMF images collected over a week after the cessation of the training set, with the dense network performing as well as the CNN over the entire period.

Optical fibers have proven to be extremely useful for endoscopy and related applications[1,2]. Present commercial methods for transmitting images through fibers are based on single-mode fiber bundles[3,4], consisting of thousands of fibers each transmitting a single pixel. It would be advantageous to instead transmit images in multimode fibers (MMFs), which are easy to fabricate and thinner than single-mode fiber bundles, and could potentially carry much more information. However, there is a serious drawback: due to mode-mixing and modal dispersion, any image coupled into a MMF is transformed into a complex speckle pattern at the output[5]. Researchers have devised various methods for reconstructing the input images from the speckle patterns, based on finding the complex transmission matrix of the MMF[6–10] or phase retrieval algorithms[11–13]. However, such methods generally require extra apparatus for measuring the optical phase, or have difficulty scaling to large image sizes.

Another promising approach is to use a training set of *a priori* known inputs to teach an artificial neural network (NN) how to map MMF output images to input images. This would not require additional interferometric equipment, and can potentially scale up to large image sizes. The idea was proposed and investigated decades ago[14–16], but only in recent years has it been shown to perform well for reconstructing images of reasonable complexity[17–21], aided by improvements in computational power and NN software.

These recent advances in NN-aided MMF image reconstruction have focused on deep convolutional neural networks (CNNs)[17–22]. Unlike traditional dense NNs[23], CNNs use convolution operations instead of general matrix multiplication within the NN layers[24], inspired by biological processes in visual perception. CNNs have enjoyed immense recent success in computer vision[25], making it natural to investigate using them for MMF image reconstruction. They have also been applied to the related problem of image reconstruction in scattering media[26–29]. However, there are grounds to question how well-suited CNNs are for analyzing speckle patterns such as those produced by MMFs, which are very different from the natural images commonly dealt with in computer vision. In MMF images, information is encoded not just locally but in the global distribution of speckles[22,30,31], whereas the localized receptive fields in convolutional layers are designed to extract relevant local features (such as edges) in natural images, rather than long-range spatial structures[32]. Traditional dense NNs can extract information from both local and global features due to the presence of the fully-connected hidden layers. Dense

[1]Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore. [2]Centre for Disruptive Photonic Technologies, Nanyang Technological University, Singapore 637371, Singapore. [3]State Key Laboratory of Quantum Optics and Quantum Optics Devices, Institute of Opto-Electronics, Shanxi University, Taiyuan 030006, China. [4]These authors contributed equally: Changyan Zhu and Eng Aik Chan. ✉email: blzhang@ntu.edu.sg; csoci@ntu.edu.sg; yidong@ntu.edu.sg

NNs have previously been employed for controlling spatiotemporal nonlinearities in MMFs[33], and for position sensing in a chaotic cavity[34].

This paper investigates the performance of dense NNs and CNNs for MMF image reconstruction. Whereas the earliest papers on NN-aided MMF image reconstruction used dense NNs[14–16], most recent studies have concentrated on using CNNs[17–22]. (One exception to this trend was the study by Turpin *et al.* of both dense NNs and CNNs for transmission control in scattering media and MMFs[28]. Another work by Caramazza *et al.* used dense NNs to approximate the transmission matrix of a MMF, as an alternative to direct image reconstruction[35]). To our knowledge, there has been no direct comparison between the two NN architectures in the context of MMF image reconstruction. The recent popularity of CNNs for this task is predicated on the local feature extraction capability of CNNs being useful for descrambling MMF images. Our comparison of dense NNs and CNNs should be useful to the community in testing this assumption.

Our principal comparison is between (i) the single hidden layer dense neural network (SHL-DNN), one of the simplest dense NN architectures, and (ii) U-Net, a CNN originally developed for biomedical imaging[36], which has recently been used for MMF image reconstruction[18]. We do not compare very deep CNNs such as Resnet[18] or generative adversarial networks[37], as these require much greater computational resources and longer training times, and thus seem ill-suited to the MMF image reconstruction problem. After optimizing both types of NNs (SHL-DNN and U-Net), we find that the SHL-DNN achieves a similarly high reconstructed image fidelity with shorter training time and less network complexity than the U-Net. For one of our reference datasets, SHL-DNN achieves a saturation Structural Similarity Index Measure (SSIM)[38] of 0.775 in 16 minutes and the U-Net achieves a saturation SSIM of 0.767 in 3.5 h on the same computer. The SHL-DNN has 20 million trainable parameters, and the U-net has 31 million trainable parameters. We also validated both NNs using images collected up to 235 h after the images in the training set; both NNs continue to perform well in image reconstruction. Moreover, we tested a "VGG-type" NN, which combines convolutional and dense layers, and found that it offers no additional performance advantage over the SHL-DNN.

Experimental setup

Multimode fiber image reconstruction. The optical setup is shown in Fig. 1a. A collimated beam from a diode laser with an operating wavelength of 808 nm (Thorlabs LP808-SF30) is expanded and directed onto a spatial light modulator (SLM) (Hamamatsu X13138-02). Along with two orthogonal polarizers, the SLM generates a programmable spatial modulation in the intensity of the light beam.

The modulated beam is coupled into a one meter long multimode fiber (MMF) (Thorlabs FT400EMT) via a matching collimator (NA 0.39). The distal end of the MMF is imaged with a CMOS camera (Thorlabs DC1545M). The camera images consist of complicated speckle patterns, as shown in the left panel of Fig. 1b, with no apparent relation to the ground truth images from the SLM. The camera images have 1280×1080 pixel resolution; to obtain a tractable dataset, we crop and downsample to 64×64 using the Lanczos algorithm[39], as shown in the right panel of Fig. 1b.

By operating the SLM with a refresh rate of 0.9 Hz (which allows for the generation of stable and distortion-free images), we accumulate one dataset of 61524 MMF images collected over approximately 19 h for training and several datasets that spans across 235 h. The ground truth images are drawn equally from (i) the MNIST digit dataset containing handwritten digits in various styles[40], and (ii) the MNIST-Fashion dataset containing images of clothing and apparel[41]. The MNIST digit dataset is used for most of the experiment; the MNIST-Fashion dataset is used in "Transfer learning and alternate image set" section.

The MNIST and MNIST-Fashion ground truth images are 28×28, whereas the MMF-derived images in the dataset are 64×64. Conceptually, there is no reason to restrict the MMF images (NN inputs) to the same size as the ground truth images (and NN outputs), as was the practice in earlier studies[17,18]. Intuitively, higher resolutions for the MMF images should be advantageous, as the image reconstruction algorithm is given more information to work with, subject to the constraints of trainability and computer memory capacity. The effects of varying the input size are studied in "Image reconstruction fidelity" section. Both the ground truth and MMF images have 8 bits of dynamic range.

Neural networks. We mainly investigate and compare two NN architectures for efficacy in MMF image reconstruction: a single hidden layer dense neural network (SHL-DNN) and the convolutional neural network U-Net. (A third architecture, a hybrid convolutional/dense network, is briefly discussed in "Hybrid neural network" section).

Dense NNs are the most elementary architecture for NN-based machine learning. The earliest papers on NN-aided MMF image reconstruction utilized dense NNs[14–16], but were constrained by the lower levels of computational power then available. We implement the SHL-DNN shown in Fig. 1c, featuring a hidden layer of 4096 nodes sandwiched between input and output layers, with dense interlayer node connections. Each 64×64 input image is flattened and inserted into the input layer, which has $64^2 = 4096$ nodes. The hidden layer and output layer have sigmoid activation functions. The result from the output layer (which has $28^2 = 784$ nodes) is reshaped into a 28×28 image that can be compared to the ground truth image.

Convolutional neural networks (CNNs) have been applied to the MMF image reconstruction problem by several recent authors[17–22]. Here, we employ the U-Net architecture, which Rahmani *et al.* have previously used for MMF image reconstruction with the MNIST digit dataset[18]. As shown in Fig. 1d, the input is 64×64 and the output is 28×28, the same as for the SHL-DNN. The network consists of a sequence of convolutional and pooling layers leading to a $4 \times 4 \times 1024$ intermediary layer, followed by a sequence of convolutional upsampling layers. A $64 \times 64 \times 64$ convolutional layer followed by a 2×2 max pooling layer downsamples the 64×64 input image to 32×32 before the $32 \times 32 \times 128$ convolutional layer of the U-Net. Batch normalization is applied after
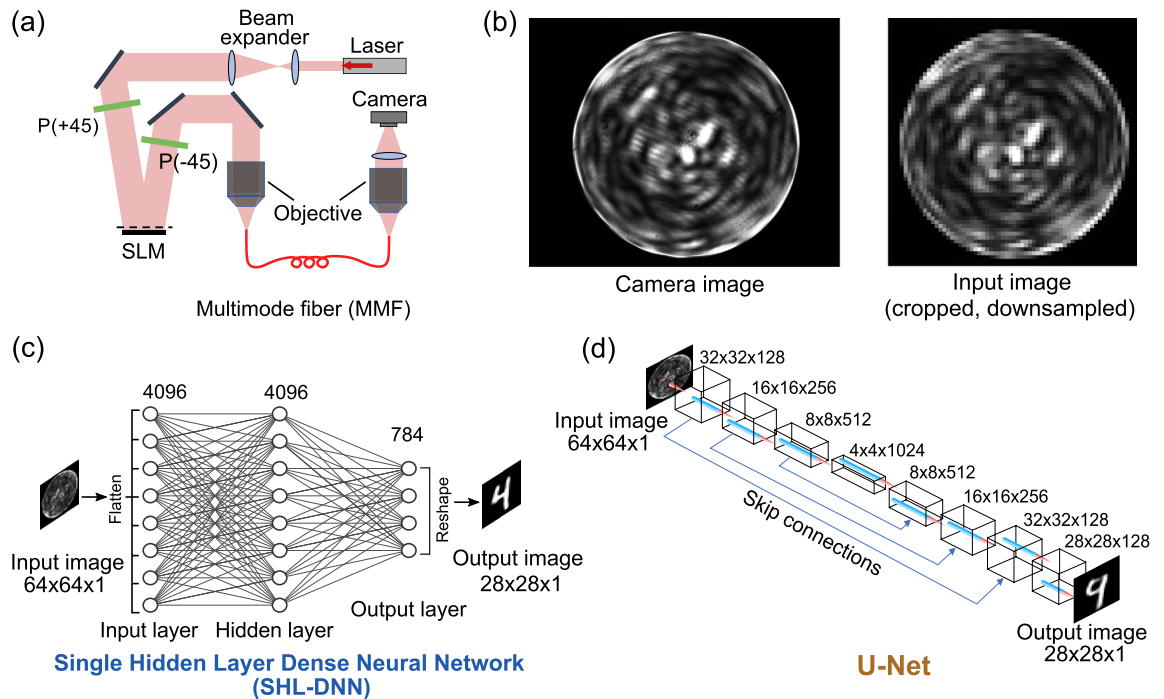
Figure 1. (**a**) Experimental setup for transmitting images through a multimode fiber. A laser beam is expanded and reflected off a spatial light modulator (SLM), which together with a pair of polarizers (P) generates an intensity modulated image. The beam is coupled into a multimode fiber (MMF), and the distal end is imaged by a camera. (**b**) Example of a scrambled image from the MMF. The ground truth image is a digit from the MNIST database (see Fig. 2). The left panel shows the full-resolution (1280 × 1080 pixels) camera image; the right panel shows the cropped and downsampled 64 × 64 image fed to the neural network. (**c**) Schematic of a single hidden layer dense neural network (SHL-DNN) with 4096 nodes in the hidden layer. The input image is flattened at the input layer, and the output is reshaped into a two-dimensional image. (**d**) Schematic of a U-Net consisting of contracting convolutional layers, an intermediary layer, and expanding convolutional layers. For each convolutional layer, the size $a \times b \times c$ refers to $a \times b$ pixels with c filters (image depth). At the input, the 64 × 64 input image is mapped onto a 64 × 64 × 64 convolution layer (not shown) before max pooling to the 32 × 32 × 128 convolutional layer. Skip connections concatenate the outputs from successive contracting layers with the corresponding expanding layers. Batch-normalization are applied to all convolutional layers.

each convolutional layer. Each convolutional layer has a ReLU activation function, and the output layer has a sigmoid activation function (similar to the SHL-DNN). We follow the typical U-Net architecture design rule[36] wherein a halving of the layer dimensions is accompanied by a doubling of the number of filters (image depth), and vice versa. There are also auxiliary skip connections that aid image localization[36].

The U-Net architecture contains numerous hyperparameters such as the number of layers, convolutional filter depths, batch size, etc. We tested the effects of varying these hyperparameters, and the "baseline" configuration shown in Fig. 1d gives the best results. Notably, in this configuration the filter depths are four times what was used in Rahmani *et al.*[18].

Each NN is trained using Adam optimization with a batch size of 256 images, and an early stopping condition of 100 epochs after validation losses stop improving. We find that batch-normalization regularization is crucial for the U-Net to perform well, but dropout regularization is better for the SHL-DNN. Little performance improvement is observed when the batch size adjusted between 128 and 1024; a much larger batch size (27685) drastically lengthens training times. For the objective function, the NN output is compared against the ground truth (MNIST digit or MNIST-Fashion) image via the Structural Similarity Index Measure (SSIM)[38], a well-established metric for quantifying the similarity between structured images (seeSupplementary Materials). All training was performed on the same computer (Intel Xeon Gold 5218 with NVIDIA Quadro RTX 5000 GPU).

Results

Image reconstruction fidelity. We train the SHL-DNN and U-Net using 30762 MMF images from the first 19 h of the data collection run. The ground truth images drawn from either the MNIST digit dataset[40] or the MNIST-Fashion dataset[41]; separate instances of each network are trained for the two respective datasets. In each case, we assign 27685 images for training and the remaining 3077 for validation. The training and validation images are initially drawn randomly from across the collection period (the role of collection time will be investigated later, in "Performance over time" section).

Figure 2 shows the results of MMF image reconstruction for six representative images from the validation set, three from the MNIST digit datasets and three from the MNIST-Fashion sets. The fully-trained SHL-DNN
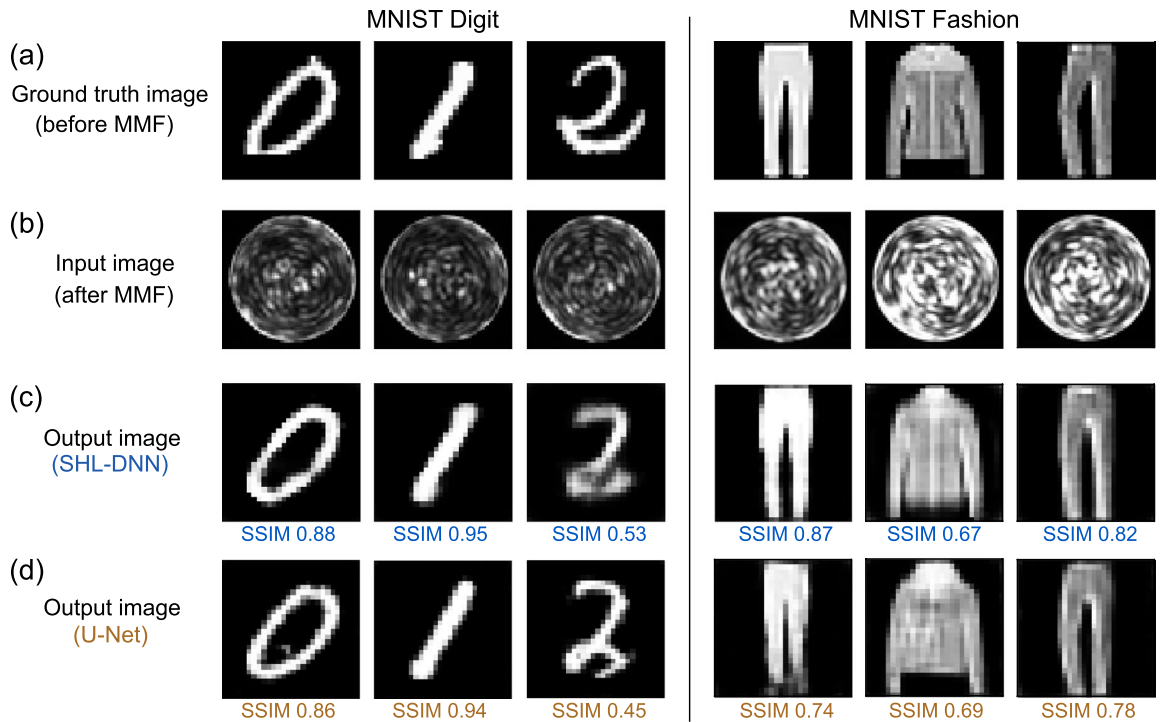
Figure 2. Demonstration of MMF image reconstruction on the MNIST digit dataset (first three columns) and the MNIST-Fashion dataset (last three columns). (**a**) A representative sample of 28 × 28 ground truth images. (**b**) The corresponding 64 × 64 images obtained from the MMF. (**c**) Reconstructed 28 × 28 images produced by the SHL-DNN. The structural similarity (SSIM) relative to the ground truth image is shown below each reconstructed image. (**d**) The corresponding results produced by the U-Net.

and U-Net both recover the ground truth images with remarkable fidelity (Fig. 2a,c–d), despite the lack of human-discernable patterns in the MMF images (Fig. 2b). The SHL-DNN and U-Net achieve similar fidelity, as corroborated by the similar SSIM scores, for both types of images.

Figure 3a shows the SHL-DNN and U-Net training curves for the MNIST digit dataset (the results for MNIST-Fashion are similar; see Supplementary Materials). We plot the training curves against elapsed time to allow for fairer comparisons, since the two networks have very different training times per epoch. The performance of the SHL-DNN saturates at SSIM 0.775, comparable to SSIM 0.767 for the U-Net. Figure 3b,c, we compare the performance of both networks on two other common metrics: the mean squared error (MSE) and the resulting classification error for the validation set (however, the training still uses SSIM for the objective function). The classification error is meant to characterize the overall legibility of the reconstructed digits, and is obtained by passing the NN outputs to an auxiliary digit classifier (`mnist_cnn.py` from Keras[42]). The results from these alternative measures are similar to what was obtained from the $\overline{\text{SSIM}}$. The SHL-DNN achieves MSE 2.25×10^{-2} and classification accuracy 0.90, while the U-Net achieves MSE 2.48×10^{-2} and classification accuracy 0.90.

Although the two networks yield similar image reconstruction fidelity, the SHL-DNN can be trained more quickly. To reach its saturation SSIM (i.e., triggering of the stopping condition), the SHL-DNN takes 462 epochs and 16 minutes, whereas the U-Net takes 318 epochs and 3.5 h. The training time per epoch is 20 times faster for the SHL-DNN.

We systematically investigated the effects of various NN settings, and found that no further major performance improvements are achievable without increasing the training set size. (For these hyperparameter studies, a smaller training set of 8709 images was utilized.) For the SHL-DNN, the choice of input image size appears to play an important role. As shown in Fig. 3d, for a smaller input image size (28 × 28) the SSIM saturates at a lower value, which can be ascribed to the NN having less information available for image reconstruction. But having inputs that are too large, such as 84 × 84, also leads to a lower SSIM compared to our baseline choice of 64 × 64. The SHL-DNN performance decreases when the number of hidden layer nodes is reduced below the baseline value, as shown by the red curve in Fig. 3d for the 512 node case. On the other hand, further increasing the number of hidden layer nodes increases the training time without significant improvement in the saturated SSIM (see Supplementary Materials). Moreover, the number of hidden layer nodes seems to have negligible influence on the optimal input image size.

As for the U-Net, one setting that notably affects performance is the number of convolutional filters. We denote the number of U-Net filters used in Rahmani et al.[18] as "1×". The saturated SSIM score increases as the number of filters in increased up to 4×, which is the baseline value that we adopted. Further increases in the number of filters leads to a substantial increase in training time, without significant performance improvement. Another possible setting is the number of convolutional layers; we verified that deeper or shallower U-Net structures adversely affect the performance. Moreover, we find that removing the skip connections leads to a
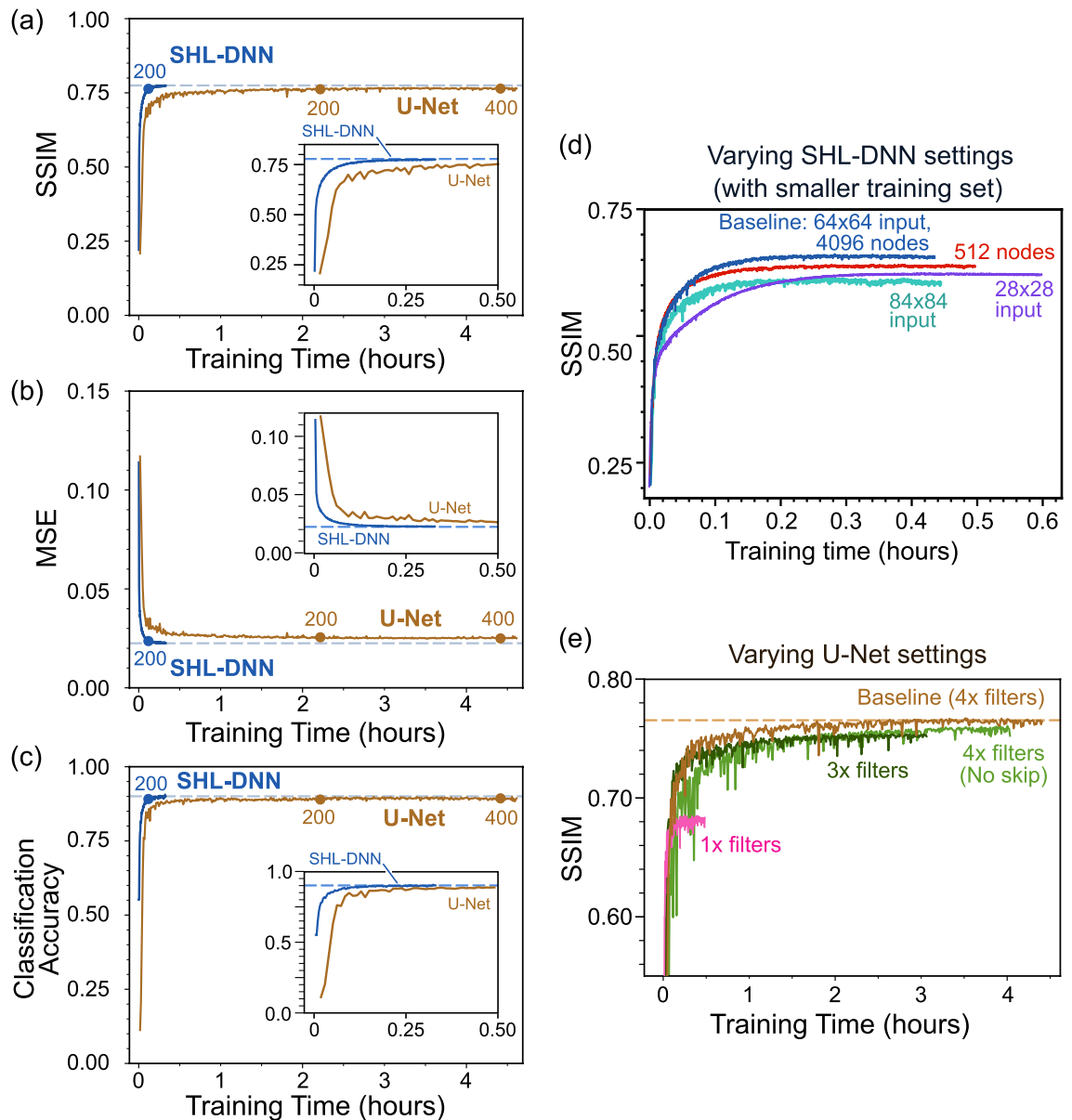
Figure 3. (**a**)–(**c**) Training curves for SHL-DNN and U-Net, using SSIM as the objective function and with 27685 training images and 3077 validation images. Epoch numbers are indicated by the numbered circles on each curve. (**a**) SSIM versus training time. (**b**) Mean squared error (MSE) versus training time. (**c**) Classification accuracy versus training time, obtained by feeding the output images from each neural network into an auxiliary classifier network, serving as a measure of legibility. In (**a**)–(**c**), the converged performance measure for the SHL-DNN is indicated by horizontal dashes. (**d**) SHL-DNN performance with different settings, calculated with a training set of 8709 images: the baseline network used in (**a**)–(**c**) and depicted in Fig. 1c, with 64×64 inputs and 4096 hidden layer nodes (blue), a network with 28×28 inputs (purple), a network with 84×84 inputs (green), and a network with 512 hidden layer nodes and 64×64 inputs (red). (**e**) U-Net performance for different settings: the baseline U-Net with "$4\times$" filters [used in (**a**)–(**c**) and depicted in Fig. 1d] (brown), removed skipping layers (light green), "$3\times$" filters (dark green), and "$1\times$" filters (pink). In the last case, convergence is achieved about as quickly as the SHL-DNN, but at significantly lower SSIM.

slight decrease in performance slightly; hence, the skip connections are included in our baseline configuration (although the reason these connections are advantageous for the present task is somewhat unclear). Some of these comparisons are shown in Fig. 3e.

After these optimization studies, we arrive at the SHL-DNN and U-Net configurations shown in Fig. 1c,d. In these configurations, the SHL-DNN has 20 million trainable parameters and takes 39.9 million FLOPs per forward pass, and the U-net has 31 million trainable parameters and takes 62.8 million FLOPs per forward pass.
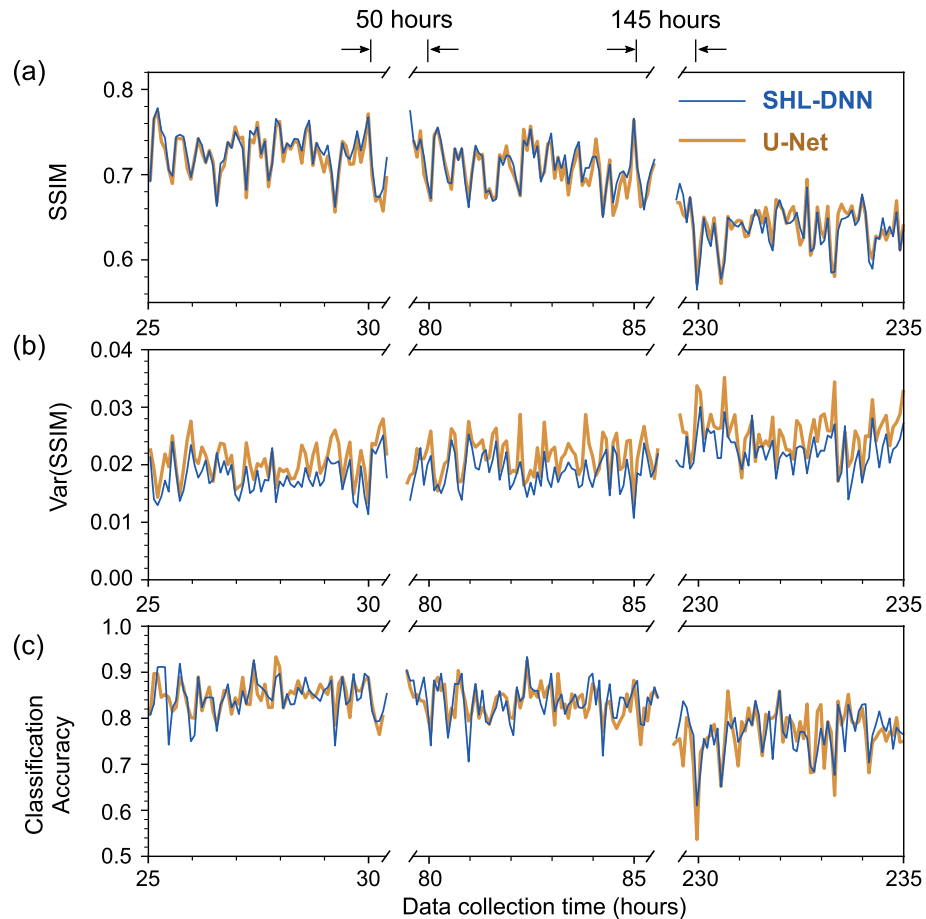
Figure 4. MMF image reconstruction metrics using data collected at different times subsequent to the training set. The SHL-DNN and U-Net are trained using 27685 images collected over 19 h, and then validated against images collected over the subsequent 235 h. The time axis is divided into 5 minute bins with 137 validation images per bin. (**a**) SSIM. (**b**) Variance of SSIM, corresponding to the spread of the SSIM in each 5 minute bin. (**c**) Classification accuracy, obtained by feeding the output images from each neural network into an auxiliary high-accuracy classifier.

Performance over time. It is interesting to ask whether the image reconstruction ability of the NNs is persistent, or whether it degrades over time due to a drift in the MMF's transmission characteristics. Such temporal changes can be caused by thermal and mechanical perturbations of the environment, which induce minute deformations of the fiber.

To address this question, we validate the NNs (trained using images from the first 19 h of the dataset) against images collected during the subsequent 235 h. The results are shown in Fig. 4. The validation data are sorted by collection time and batched into 5 minute intervals.

In terms of both SSIM and digit classification accuracy, the image reconstruction performance for both NNs fluctuates over time, but is overall remarkably robust. It can be noticed in Fig. 4a,c that the performance fluctuations for the SHL-DNN and U-Net are correlated over time. In fact, their SSIM scores have a correlation coefficient of 0.950. This implies that the performance fluctuations are caused by the MMF undergoing physical fluctuations in its transmission characteristics (relative to the training set), which simultaneously impacts the performance of both NNs. Over the 235 h period, we observe only a slight long-term degradation in performance (both in terms of SSIM and digit classification accuracy), indicating that there is negligible sustained "drift" in the MMF's transmission characteristics. Over the entire experimental period, the SHL-DNN and U-Net consistently have similar performance, with SSIM variance of about 0.02.

Hybrid neural network. Rahmani et al.[18] studied the use of another type of NN for unscrambling MMF images: a hybrid convolutional and dense network of the type pioneered by Oxford's Visual Geometry Group (VGG). VGG-type networks are typically used for classification[43], and they were used in Ref.[18] for digit classification with the MNIST digit dataset. In this paper, we are mainly interested in image *reconstruction* rather than *classification*. Nonetheless, it is helpful to study the performance of a VGG-type network for this purpose, as a further test of the usefulness of convolutional layers for extracting structural information from MMF images.
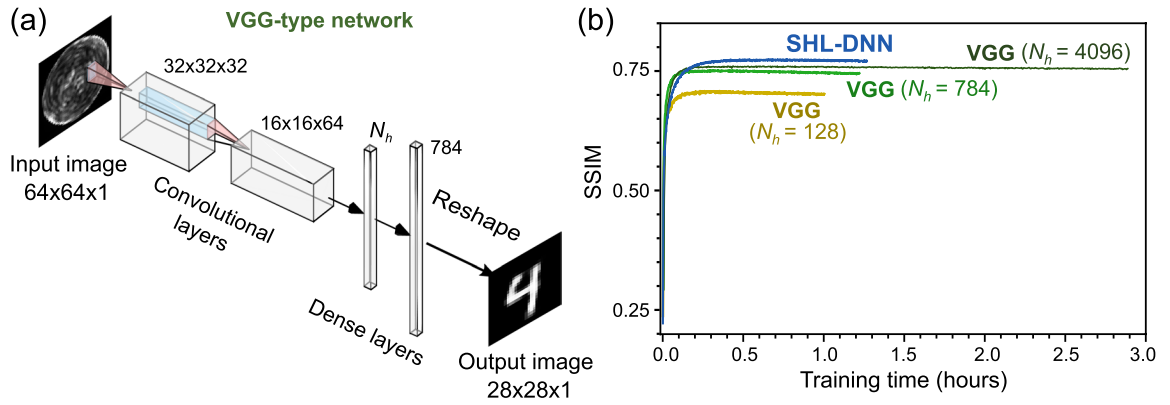
Figure 5. Performance of a VGG-type network for MMF image reconstruction. (**a**) Schematic of the VGG-type network, which consists of two convolutional layers, a dense hidden layer with N_h nodes, and a dense output layer. (**b**) Training curves for SHL-DNN and VGG-type networks: the baseline SHL-DNN corresponding to Fig. 1c, with 4096 hidden layer nodes (blue), and VGG-type networks with $N_h = 4096$ (dark green), $N_h = 784$ nodes (light green), and $N_h = 128$ (yellow).
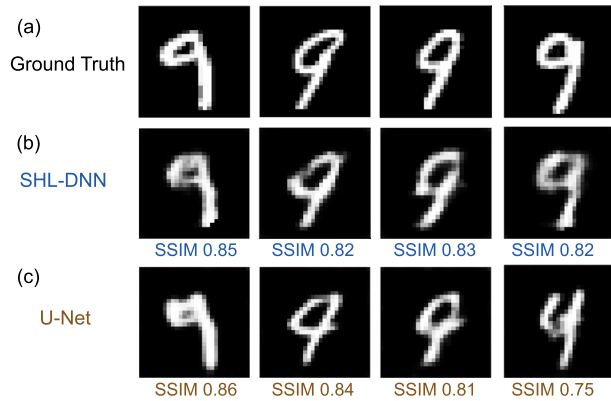
Figure 6. Reconstruction of images of the digit 9 from the MNIST digit dataset, using a SHL-DNN trained with a modified MNIST digit dataset excluding all instances of the digit 9. (**a**) Ground truth images. (**b**) Reconstructed images of digit '9' by the SHL-DNN trained with the modified dataset. (**c**) Reconstructed images of digit '9' by the U-net trained with the modified dataset. SSIM scores are shown below the reconstructed images.

We implement a simple VGG-type network as shown in Fig. 5a, consisting of two convolutional layers, a hidden dense layer with N_h nodes, and a dense output layer. Figure 5b shows the training curves for VGG-type networks with several choices of N_h, as well as for the baseline SHL-DNN. When N_h is equal to the number of hidden layer nodes in the SHL-DNN, the saturated SSIM is 0.71—comparable to but certainly not better than the SHL-DNN (SSIM 0.775). For smaller values of N_h, the performance is substantially worse. We also investigated reversing the configuration by placing the dense layers at the input and the convolutional layers at the output, but this not produce any improvement. These results seem to bolster the case that convolutional input layers do not provide additional benefits for MMF image reconstruction, a point that will be further discussed in "Discussion" section.

Transfer learning and alternate image set. Transferability is a common concern in machine learning. In the present context, one may ask whether NNs trained using one kind of ground truth image—say, MNIST digits—can successfully reconstruct more general images. In other words, are the networks broadly capable of undoing the effects of mode mixing in the MMF, or are they merely recognizing patterns that are highly specific to the sort of images in the training set?

To investigate this, we train the SHL-DNN by withholding one digit from the MNIST digit dataset, and validating it against the omitted digit. Figure 6a shows representative results for the case of an omitted digit '9'. Although this SHL-DNN has not seen any examples based on the digit '9', it reconstructs the images reasonably well, albeit with lower SSIM, as shown in Fig. 6b. Here, the training set (with '9' excluded) has 14565 images, and the other network settings are the same as in the baseline network described in "Image reconstruction fidelity" section. Over 1000 instances of the digit '9', the mean SSIM is 0.72, compared to SSIM 0.86 for a validation set

7

of 2913 images that exclude the digit '9'. The performance of U-net is quite similar to SHL-DNN, as shown in Fig. 6c: the mean SSIM is 0.70 over 1000 instances of the digit '9'.

When we attempt to reconstruct MNIST-Fashion images using a SHL-DNN trained on MNIST digits, or vice versa, the results are extremely poor (SSIM close to zero). Likewise, when we attempt to reconstruct images consisting of random uncorrelated pixel intensities, all three trained networks (SHL-DNN, U-Net, and VGG-type) give very poor results; over 1000 images, the MSE is in the range of 0.08–0.09 for all the three networks, comparable to the nascent training stage of Fig. 3b.

Discussion

We find that CNNs offer no performance advantage over the traditional dense NN architecture for MMF image reconstruction. In fact, the tested SHL-DNN is able to produce the same results as U-Net with much shorter training time and less network complexity, and this seems to be robust over various different NN settings. Moreover, a VGG-type hybrid convolutional/dense NN offers no obvious improvement over the SHL-DNN. These results suggest that the SHL-DNN, though a simple architecture, already gives a ceiling for the performance of neural networks on MMF image reconstruction. For practical real-time imaging applications, simpler NN architectures may be desirable as they can be trained more quickly and with fewer computational resources.

Our interpretation of the situation is that convolutional layers, though well-suited to extracting local features in natural images, do not provide any special advantage in processing speckle patterns of the sort produced by MMFs[30,31]. These speckle patterns are known to contain global correlations created by the superposition and dispersion of different transmission modes in the MMF[44,45]. By design, CNNs excel at processing local features and do not perform as well in processing global features[46]. It would be interesting to explore modifications to the CNN scheme, or preprocessing schemes for the speckle pattern, to improve performance[22].

The trained NNs can reliably reconstruct images collected long after the training set. Specifically, after training over a period of 19 h, robust image reconstruction can be achieved over the subsequent 235 h, with no degradation corresponding to a long-term drift in the fiber transmission characteristics. The implication is that the 19 h training is sufficiently representative of the perturbations (thermal fluctuations and weak dynamic disturbances) that can occur during the validation period[19,47]. Fan *et al.* have suggested that NNs can even be trained to perform image reconstruction over large varieties of deformations and configurations, such as twisting the MMF in different ways[19].

The NNs perform poorly on images that are too different from those in the training set, which is a common problem with NN-based machine learning. Recently, Caramazza *et al.* have demonstrated using an optimization algorithm to learn the complex transmission matrix for MMF image reconstruction[35], which bypasses the transfer learning limitations of the NN approach. However, this method requires much more computer memory, and the resulting image fidelity is lower; from our testing based on the same dataset from MNIST digits, the SSIM scores are ~ 0.42, compared to ~ 0.75 for the SHL-DNN. In the future, it would be interesting to attempt to combine these two approaches in a way that overcomes their individual limitations.

References

1. Seibel, E. J., Johnston, R. S. & Melville, C. D. A full-color scanning fiber endoscope. In Gannot, I. (ed.) *Optical Fibers and Sensors for Medical Diagnostics and Treatment Applications VI*, vol. 6083, 9 – 16, https://doi.org/10.1117/12.648030. International Society for Optics and Photonics (SPIE, 2006).
2. Lee, C., Engelbrecht, C., Soper, T., Helmchen, F. & Seibel, E. Scanning fiber endoscopy with highly flexible, 1 mm catheterscopes for wide-field, full-color imaging. *J. Biophoton.* **3**, 385–407. https://doi.org/10.1002/jbio.200900087 (2010).
3. Porat, A. *et al.* Widefield lensless imaging through a fiber bundle via speckle correlations. *Opt. Express* **24**, 16835–16855. https://doi.org/10.1364/OE.24.016835 (2016).
4. Shinde, A., Perinchery, S. M. & Murukeshan, V. M. A targeted illumination optical fiber probe for high resolution fluorescence imaging and optical switching. *Sci. Rep.* **7**, 45654–45654. https://doi.org/10.1038/srep45654 (2017).
5. Gover, A., Lee, C. P. & Yariv, A. Direct transmission of pictorial information in multimode optical fibers. *J. Opt. Soc. Am.* **66**, 306–311. https://doi.org/10.1364/JOSA.66.000306 (1976).
6. Choi, Y. *et al.* Scanner-free and wide-field endoscopic imaging by using a single multimode optical fiber. *Phys. Rev. Lett.* **109**, 203901 (2012).
7. Caravaca-Aguirre, A. M., Niv, E., Conkey, D. B. & Piestun, R. Real-time resilient focusing through a bending multimode fiber. *Opt. Express* **21**, 12881–12887 (2013).
8. Gu, R. Y., Mahalati, R. N. & Kahn, J. M. Design of flexible multi-mode fiber endoscope. *Opt. Express* **23**, 26905–26918 (2015).
9. Loterie, D. *et al.* Digital confocal microscopy through a multimode fiber. *Opt. Express* **23**, 23845–23858 (2015).
10. Rothe, S., Radner, H., Koukourakis, N. & Czarske, J. W. Transmission matrix measurement of multimode optical fibers by mode-selective excitation using one spatial light modulator. *Appl. Sci.* **9**, 195 (2019).
11. Popoff, S. *et al.* 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).
12. N'Gom, M., Estakhri, N., Norris, T. B., Michielssen, E. & Nadakuditi, R. R. Controlled transmission through highly scattering media using semi-definite programming as a phase retrieval computation method. In *2018 Conference on Lasers and Electro-Optics (CLEO)*, 1–2 (IEEE, 2018).
13. N'Gom, M., Norris, T. B., Michielssen, E. & Nadakuditi, R. R. Mode control in a multimode fiber through acquiring its transmission matrix from a reference-less optical system. *Opt. Lett.* **43**, 419–422 (2018).
14. Aisawa, S., Noguchi, K. & Matsumoto, T. Remote image classification through multimode optical fiber using a neural network. *Opt. Lett.* **16**, 645–647 (1991).
15. Matsumoto, T., Koga, M., Noguchi, K. & Aizawa, S. Proposal for neural-network applications to fiber-optic transmission. In *1990 IJCNN International Joint Conference on Neural Networks*, 75–80 (IEEE, 1990).

16. Marusarz, R. K. & Sayeh, M. R. Neural network-based multimode fiber-optic information transmission. *Appl. Opt.* **40**, 219–227 (2001).
17. Borhani, N., Kakkava, E., Moser, C. & Psaltis, D. Learning to see through multimode fibers. *Optica* **5**, 960–966 (2018).
18. Rahmani, B., Loterie, D., Konstantinou, G., Psaltis, D. & Moser, C. Multimode optical fiber transmission with a deep learning network. *Light Sci. Appl.* **7**, 1–11 (2018).
19. Fan, P., Zhao, T. & Su, L. Deep learning the high variability and randomness inside multimode fibers. *Opt. Express* **27**, 20241–20258 (2019).
20. Yang, M. *et al.* Deep hybrid scattering image learning. *J. Phys. D Appl. Phys.* **52**, 115105 (2019).
21. Rahmani, B. *et al.* Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media. *Nat. Mach. Intell.* **2**, 403–410 (2020).
22. Kakkava, E. *et al.* Imaging through multimode fibers using deep learning: the effects of intensity versus holographic recording of the speckle pattern. *Opt. Fiber Tech.* **52**, 101985. https://doi.org/10.1016/j.yofte.2019.101985 (2019).
23. Braspenning, P. J., Thuijsman, F. & Weijters, A. J. M. M. *Artificial Neural Networks: An Introduction to ANN Theory and Practice* Vol. 931 (Springer, New York, 1995).
24. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, Cambridge, 2016).
25. Rawat, W. & Wang, Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29**, 2352–2449. https://doi.org/10.1162/neco_a_00990 (2017).
26. Li, Y., Xue, Y. & Tian, L. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica* **5**, 1181–1190 (2018).
27. Li, S., Deng, M., Lee, J., Sinha, A. & Barbastathis, G. Imaging through glass diffusers using densely connected convolutional networks. *Optica* **5**, 803–813 (2018).
28. Turpin, A., Vishniakou, I. & d Seelig, J. Light scattering control in transmission and reflection with neural networks. *Opt. Express* **26**, 30911–30929. https://doi.org/10.1364/OE.26.030911 (2018).
29. Sun, Y., Shi, J., Sun, L., Fan, J. & Zeng, G. Image reconstruction through dynamic scattering media based on deep learning. *Opt. express* **27**, 16032–16046 (2019).
30. Redding, B. & Cao, H. Using a multimode fiber as a high-resolution, low-loss spectrometer. *Opt. Lett.* **37**, 3384–3386. https://doi.org/10.1364/OL.37.003384 (2012).
31. Pauwels, J., Van der Sande, G. & Verschaffelt, G. Space division multiplexing in standard multi-mode optical fibers based on speckle pattern classification. *Sci. Rep.* **9**, 17597. https://doi.org/10.1038/s41598-019-53530-6 (2019).
32. Linsley, D., Kim, J., Veerabadran, V., Windolf, C. & Serre, T. Learning long-range spatial dependencies with horizontal gated recurrent units. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 152 (ACM, 2018).
33. Teğin, U. *et al.* Controlling spatiotemporal nonlinearities in multimode fibers with deep neural networks. *APL Photon.* **5**, 030804 (2020).
34. del Hougne, P. Robust position sensing with wave fingerprints in dynamic complex propagation environments. *Phys. Rev. Res.* **2**, 043224 (2020).
35. Caramazza, P., Moran, O., Murray-Smith, R. & Faccio, D. Transmission of natural scene images through a multimode fibre. *Nat. Commun.* **10**, 1–6 (2019).
36. Ronneberger, O., Fischer, P. & Brox, T. U-net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241 (Springer, 2015).
37. Zhang, X. *et al.* Experimental demonstration of a multimode fiber imaging system based on generative adversarial networks. In *Asia Communications and Photonics Conference (ACPC) 2019*, T4A.4 (Optical Society of America, 2019).
38. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
39. Madhukar, B. & Narendra, R. Lanczos resampling for the digital processing of remotely sensed images. In *Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN-2013)*, 403–411 (Springer India, India, 2013).
40. LeCun, Y., Cortes, C. & Burges, C. Mnist handwritten digit database. *ATT Labs [Online]*. Available: http://yann.lecun.com/exdb/mnist **2** (2010).
41. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms (2017). arXiv:1708.07747.
42. Chollet, F. *et al.* Keras. https://github.com/fchollet/keras (2015).
43. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition (2014). arXiv:1409.1556.
44. Xiong, W., Hsu, C. W. & Cao, H. Long-range spatio-temporal correlations in multimode fibers for pulse delivery. *Nat. Commun.* **10**, 1–7 (2019).
45. Dandliker, R., Bertholds, A. & Maystre, F. How modal noise in multimode fibers depends on source spectrum and fiber dispersion. *J. Lightw. Technol.* **3**, 7–12 (1985).
46. Zheng, Y., Huang, J., Chen, T., Ou, Y. & Zhou, W. Processing global and local features in convolutional neural network (cnn) and primate visual systems. In *Mobile Multimedia/Image Processing, Security, and Applications 2018*, vol. 10668, 1066809 (International Society for Optics and Photonics, 2018).
47. Flaes, D. E. B. *et al.* Robustness of light-transport processes to bending deformations in graded-index multimode waveguides. *Phys. Rev. Lett.* **120**, 233901 (2018).

Author contributions

E.A.C., C.Z., C.S., and Y.C. conceived the project, E.A.C., W.P., and Y.W. set up and performed the experiment, and C.Z. and E.A.C. did the neural network training. All authors contributed to the writing of the manuscript. Y.C., C.S., and B.Z. supervised the project. All authors discussed the results and contributed substantially to the work.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79646-8.

Correspondence and requests for materials should be addressed to B.Z., C.S. or Y.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.