Low-power multimode-fiber projector outperforms shallow-neural-network classifiers

Daniele Ancora,[1,3,5,*] Matteo Negri,[1,3] Antonio Gianfrate,[2] Dimitris Trypogeorgos,[2] Lorenzo Dominici,[2] Daniele Sanvitto,[2] Federico Ricci-Tersenghi,[1,3,4] and Luca Leuzzi[3,1]

[1] *Department of Physics, Università di Roma la Sapienza, Piazzale Aldo Moro 5, Rome I-00185, Italy*
[2] *Institute of Nanotechnology, Consiglio Nazionale delle Ricerche (CNR-NANOTEC), Via Monteroni, Lecce I-73100, Italy*
[3] *Institute of Nanotechnology, Soft and Living Matter Laboratory, Consiglio Nazionale delle Ricerche (CNR-NANOTEC), Piazzale Aldo Moro 5, Rome I-00185, Italy*
[4] *Istituto Nazionale di Fisica Nucleare, sezione di Roma1, Piazzale Aldo Moro 5, Rome I-00185, Italy*
[5] *Epigenetics and Neurobiology Unit, European Molecular Biology Laboratory (EMBL Rome), Via Ramarini 32, Monterotondo 00015, Italy*

(Received 26 May 2023; revised 18 December 2023; accepted 1 May 2024; published 12 June 2024)

In the domain of disordered photonics, the characterization of optically opaque materials for light manipulation and imaging is a primary aim. Among various complex devices, multimode optical fibers stand out as cost-effective and easy-to-handle tools, making them attractive for several tasks. In this context, we use these fibers as random hardware projectors, transforming an input dataset into a higher-dimensional speckled image set. The goal of our study is to demonstrate that using such randomized data for classification by training a single logistic regression layer improves accuracy compared to training on direct raw images. Interestingly, we found that the classification accuracy achieved is higher than that obtained with the standard transmission-matrix model, a widely accepted tool for describing light transmission through disordered devices. We conjecture that this improved performance could be due to the hardware classifier operating in a flatter region of the loss landscape when trained on fiber data, which aligns with the current theory of deep neural networks. These findings suggest that the class of random projections operated by multimode fibers generalize better to previously unseen data, positioning them as promising tools for optically assisted neural networks. With this study, we seek to contribute to advancing the knowledge and practical utilization of these versatile instruments, which may play a significant role in shaping the future of neuromorphic machine learning.

DOI: 10.1103/PhysRevApplied.21.064027

I. INTRODUCTION

There is currently no sound understanding of the enormous success of neural networks (NNs) in learning processes and inference tasks. There is a fundamental need to understand why such architectures, which can have billions of parameters, do not severely overfit data, as predicted by statistical learning theory and the so-called *bias-variance trade-off* (see for example Refs. [1] and [2]). The abundance of learnable parameters, in fact, is arguably the most universal feature in the zoo of NN architectures. Interestingly, it is known that, given a chosen NN architecture, most of the model parameters adapt little or not at all during the learning procedure [3,4], suggesting that random projections may play an equally important role in NNs. Recent works, in fact, have shown that it is possible to train a simple two-layer model by learning only the upper layer, interpreting the first one as a random projection [5,6]. These results were strengthened further by Baldassi et al. [7], who demonstrated that increasing the dimension of the random projection leads to the production of wide and flat regions in the loss landscape (the function that is minimized during the training of the model), which are related to good generalization properties in neural networks. The generalization ability of a neural network that has been trained over a given dataset (*training dataset*) refers to its ability to display good performance when applied to data over which it was not trained (*test dataset*). In the framework of the loss-landscape description, an improvement in the generalization ability means that models that lie in flat regions make fewer mistakes when they classify previously unseen data. Finally, recent evidence has been provided [8] indicating that the way the

[*] Corresponding author: daniele.ancora@uniroma1.it, daniele.ancora@cnr.it

random projection is chosen is fundamental to determining the generalization properties of the upper layer of these simple models. This suggests that different classes of random (possibly nonlinear) projections impact differently on the performance of the models.

In this context, we are interested in studying hardware random projectors, such as those employed in the field of photonic neuromorphic computing [9,10]. The advantage of using optical neural networks (ONNs) is that neurons can interact by exploiting light scattering [11–13] and photon interference [14,15] at the speed of light. Tools for shaping and controlling the light field [16] are becoming so versatile that the discipline is under constant development, aiming at high-speed, high-throughput optical-based computing architectures. All-optical neural networks [11,17] in particular have the potential to be great tools for fast computation, though they often require an accurate model of the optical system to perform consistent back-propagation updates [18]; however, fine-tuning of the optical parameters is challenging due to discrepancies between the response of the real system and the physical model employed to describe the architecture. This *reality gap* often reduces the expected performance of the network [19,20], requiring additional corrections at the software level [12], training enforcement via hybrid strategies [18], or the use of NNs to more accurately model the optical response of the system [20].

In this rapidly evolving scenario, the class of random projections realized by multimode fibers (MMF) are promising candidates for developing ONNs. These devices scramble the photons due to scattering events occurring during light-field propagation, yielding to the formation of speckle patterns that are, in fact, random projections. Although the light transmission can be regarded as a linear process [21] in which input modes are coupled with output modes via a complex transmission rule, interference takes place when dealing with the measurement of the light-field intensity. Since the detection is nonlinear, MMFs can be used [22] to classify time-domain waveforms (using saturation effects as further nonlinearity) [23], in pattern classification of two-bit sequences [24], or for binary (human/not human) facial recognition [25]. Furthermore, when dealing with more complex classification tasks, high-power laser pulses have been employed to trigger the nonlinear response of the fiber itself [26]. Due to the increasing interest in the employment of MMFs as random-projector computing devices, we decided to study their behavior in carrying out classification tasks in a linear, low-power continuum regime. Although our MMF-based optical neural network does not employ feedback, we will show how its classification performance is considerable, as in reservoir computing systems [27–30].

We do this by comparing the performance of the physical neural network to that obtained with random Gaussian linear projections and to that of a transmission-matrix approach, which is the model commonly used to describe light propagation in disordered structures [21,31]. We performed our study statistically, shuffling the training set to assess the average behavior of the optical computing under different training and initialization conditions. Remarkably, a single MMF simultaneously provides two independent (though deterministically linked) projections, at either end of the fiber, which we studied separately using different saturation regimes. Here, we show that the real physical MMF leads to higher accuracy than its corresponding transmission-matrix model, highlighting the *reality gap* between model theory and experimental results. To assess the reason for this performance gap, we study the characteristics of complex-valued random projections in terms of the flatness of the local energy landscape, demonstrating that the MMF projection is more robust than those provided by alternative datasets. Additionally, we characterize the behavior of a hardware-based neural network using optical fibers in terms of the numbers of modes employed. We set up our study not to achieve the best performance in classification tasks, but rather to deepen the understanding of physical neural networks against their physical models, giving insights into the use of MMFs for optical computation.

II. MATERIALS AND METHODS

In a low-power regime, a generic multimode fiber transports the electromagnetic field via a linear process [21] such that the light propagation can be described using a simple multiplication of the input signal by a matrix that encodes the transmission rule:

$$\mathbf{y} = \mathbb{T}\mathbf{x}. \tag{1}$$

In this descriptive model, $\mathbf{x}$ is the controlled input, $\mathbb{T}$ is the (complex-valued and typically unknown) transmission matrix of the medium, and $\mathbf{y}$ is the output field. Despite its propagation, the way we measure the MMF output is not linear for two reasons. First, photons carry complex signals; i.e., the electromagnetic field associated with each propagation mode is characterized by amplitude and phase. Current electronic devices cannot follow the rapid oscillation of the field, which makes measurement of the phase information impossible. Assuming the possibility that the readout is also perturbed by additive noise ε, the camera will only see the noise-affected intensity distribution:

$$|\mathbf{y}|^2 = |\mathbb{T}\mathbf{x}|^2 + \varepsilon. \tag{2}$$

Second, the camera has a well-defined sensitivity range that depends on each pixel's ability to store intensity changes. If the signal reaching a given pixel exceeds its sensitivity range, the measurement gets clipped at peaks (overexposure) or at the lowest values (underexposure). In analogy to machine-learning terminology, the measurement process can be described by a nonlinear *activation function* $\sigma(\cdot)$ that acts on the result of a complex-valued linear transmission, $\mathbb{T}\mathbf{x}$. For instance, the camera's recording process can be represented using the saturating linear transfer function (satlin):

$$\sigma(\mathbb{T}\mathbf{x}) = \min\left(\max\left(d, |\mathbb{T}\mathbf{x}|^2 + \varepsilon\right), 2^b - 1\right), \qquad (3)$$

where the quantity d is the intensity threshold under which the measure is not recorded and b is the bit depth of the camera.

These considerations make the readout of a coherent field nonlinear, along with its inverse transmission recovery problem [21,31–36]. Such a matrix can be estimated using the four-phases method [21], Bayesian optimization [37], or iterative Gerchberg-Saxton schemes [38,39]; however, the characterization of the device in terms of its transmission rule is not the main scope of this paper, nor is circumventing the limitations of the measuring process. Instead, we want to study the multimodal random-projection nature of the fiber to perform optical computing. In the neural-network framework, the fiber can be seen as an optical analogy of a densely connected network composed of a single "hidden layer" with fixed weights [40]. In this shallow architecture, the MMF layer already contains a particular realization of static weights (the transmission matrix $\mathbb{T}$), which depends upon the physical status of the optical fiber. This property allows random but deterministic projections to be performed at the speed of light using a fixed transmission rule, which can be read out by the camera. Given these considerations, MMF is a good candidate for performing nonlinear optical computation using a continuous laser source, even using inexpensive and large (thus easier to handle in a setup) optical multimode fibers. In particular, if we let just a few modes propagate into the input facet of an MMF that supports many more, all the output modes will be activated, implying a few-to-many mapping. In this latter case, the optical hidden layer (i.e., MMF and camera) can perform densely connected random projections on a higher-dimensional space.

The goal of this study is to carry out image classification by concatenating a software-trained linear layer and the measured output from an MMF, as produced by inserting a given image from a dataset into the input edge of the fiber [see Fig. 1(a)]. We choose to approach the Modified National Institute of Standards and Technology (MNIST) classification problem to carry out a widely studied nonlinear task. The only parameters that we train are those of a simple logistic regression layer, which is known to achieve poor performance on the standard MNIST dataset, reaching a maximum classification accuracy of 92.7% [29]. Exploiting the random projection provided by the MMF, an optical device that is known to be linear, we compare the results with the performances obtained using reference datasets. In this study, we train the parameters of the logistic classifier using six different input datasets:

(1) *Original MNIST.* The standard MNIST dataset, comprising images of $l \times l$ pixels. The accuracy performance of this set is the baseline of our study.

(2) *Upscaled MNIST.* Each image at the original resolution is expanded by a factor L/l using a linear spline interpolation to reach the target size of $L \times L$.

(3) *Randomized MNIST.* The MNIST dataset is linearly multiplied by a Gaussian random matrix with positive entries. This maps the dataset into a higher-dimensional space, producing images with a side $L \gg l$ pixels.

(4) *MMF α-cam.* The speckled output of the MMF is recorded with a resolution of $L \times L$ pixels using camera α. Each speckle pattern is the result of sending a MNIST image on the input edge of the fiber and recording the output after disordered propagation. The patterns in the input are intensity-modulated in real space, and they have a size of $l \times l$.

(5) *MMF β-cam.* The same as MMF α-cam, with the speckles being recorded on the same input facet as that of the light injection using camera β. A relatively small portion of the light propagating forward is internally reflected and comes back toward the input edge. This determines a different speckle realization, which we acquire as an independent measurement.

(6) *MMF α-simulated.* The transmission is characterized retrieving its corresponding matrix $\mathbb{T}$ using the SmoothGS protocol [39]. The inferred transmission is used to simulate the propagation of the MNIST dataset using Eq. (1), recording the simulated speckle pattern by storing only the squared modulus.

All the datasets are used for supervised training, in which each image of the MNIST dataset is associated with the number that it represents, and the speckle image is associated with the classified number corresponding to the MNIST image impinging onto the fiber. Further details of the training procedure can be found in Appendix A. To isolate any possible dependence on the problem size, we choose to set the size of the randomized and upscaled MNIST sets to have the same dimension as the recorded fiber output. This implies that the same number of parameters are trained while solving the classification problem for every dataset, the only exception being the original set.

III. RESULTS AND DISCUSSION

In the following, we report the average results obtained by running independent logistic regressions on each

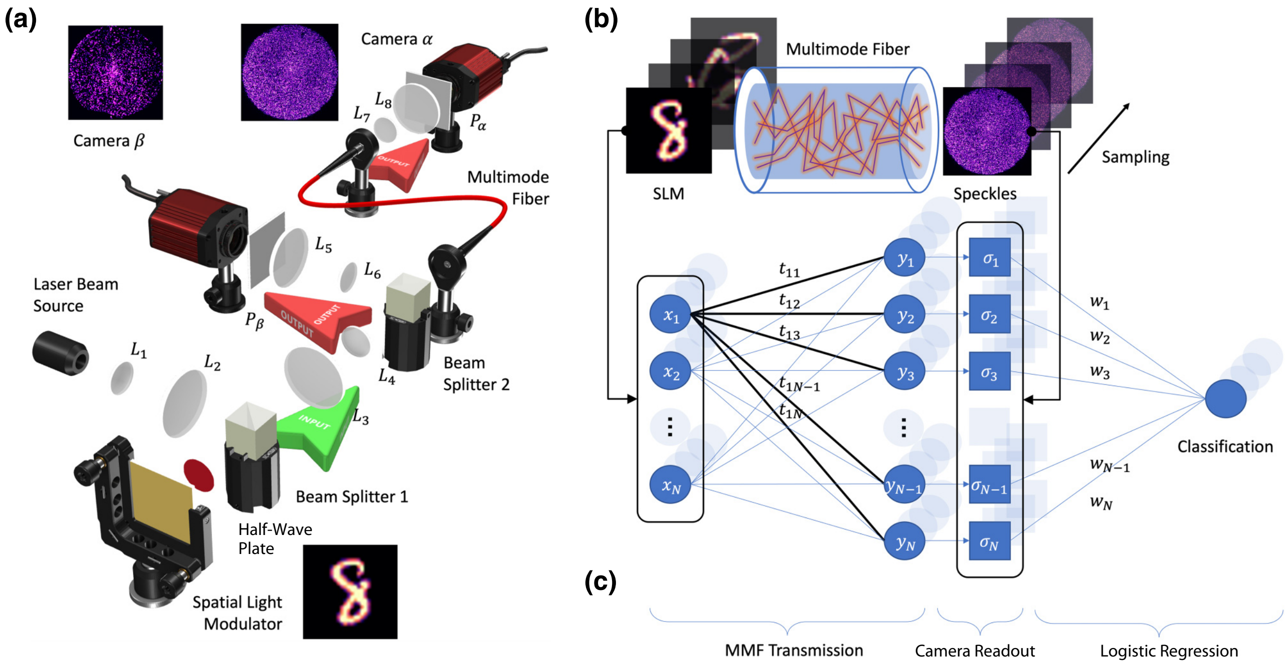

FIG. 1. Schematics of the shallow optical neural network with the MMF. (a) Simplified scheme of light transport through multimode fibers. The MNIST dataset, modulated by a spatial light modulator (SLM), enters the MMF on the input facet. During propagation, the light gets scrambled by a random but deterministic process, giving rise to the speckle pattern measured by the camera. (b) Corresponding neural-network interpretation of the light-propagation scheme. The MNIST dataset constitutes the input vector of a linear complex layer with static weights. The nonlinear operation is determined by the camera that reads the intensity of a complex field. Successively, a linear classification layer is trained using the output of the fiber. (c) Scheme of the imaging setup.

dataset, comparing the classification accuracy on a test set comprising 1000 numbers isolated from the original one.

A. Performance of different classes of projector

In experiment 1, we use 10^4 MNIST images, randomly picking up to 9000 images for training and 1000 images for testing, using $L = 600$. We repeat the parameter optimization a total of $T = 100$ times, varying the number of training samples for statistical purposes. To test the robustness of our results after training, we compute the test accuracy, which is the fraction of correctly classified data points in the test set. From Fig. 2(a), we can see how the performances of the MNIST dataset (original, randomized, and upscaled) are similar to one another. The classification problem, in fact, is well known to be a nonlinear task, and it barely generalizes using a linear model alone. Instead, using the MMF, higher performances are achieved, approaching 96% test accuracy on average on the largest set used (9000 training samples). We stress that this accuracy is not high in absolute terms because deep neural networks with convolutional layers have been able to reach more than 99% test accuracy on MNIST [41], with modern deep architectures even reaching as high as 99.91% [42]. However, we are interested in the study of the simplest ONN architecture, consisting only of a hardware random-projector layer followed by a linear classifier. With this straightforward setup, the MMF permits substantial improvement of the results obtained against a plain linear classifier (88% accuracy with 9000 training samples).

We point out that we did not use the entire MNIST dataset (composed of 60 000 images for training and 10 000 for testing) but a fraction of it; the plot trend in Fig. 2 suggests that there is room for further improvement by increasing the number of training samples. After only around 500 samplings, the gain provided by the ONN approach starts to become evident, and with only 9000 images, we can achieve performance hitting approximately 97%. To achieve the highest accuracy with the experimental data [blue and orange dots in the plot of Fig. 2(a)], we tested 100 independently initialized optimizations. Interestingly, the performance is independent of the microscopic MMF arrangement, as the two different transmission rules determined by the α and β detections perform identically. As a final note, we decided not to tune the hyperparameters of the classifier, so we can expect that their careful selection (mainly the $l2$-regularization strength and the stopping threshold) could improve the accuracy curves for all the datasets. In fact, we are not interested in the absolute numbers: our scope is to highlight the improvement determined by the physics of the

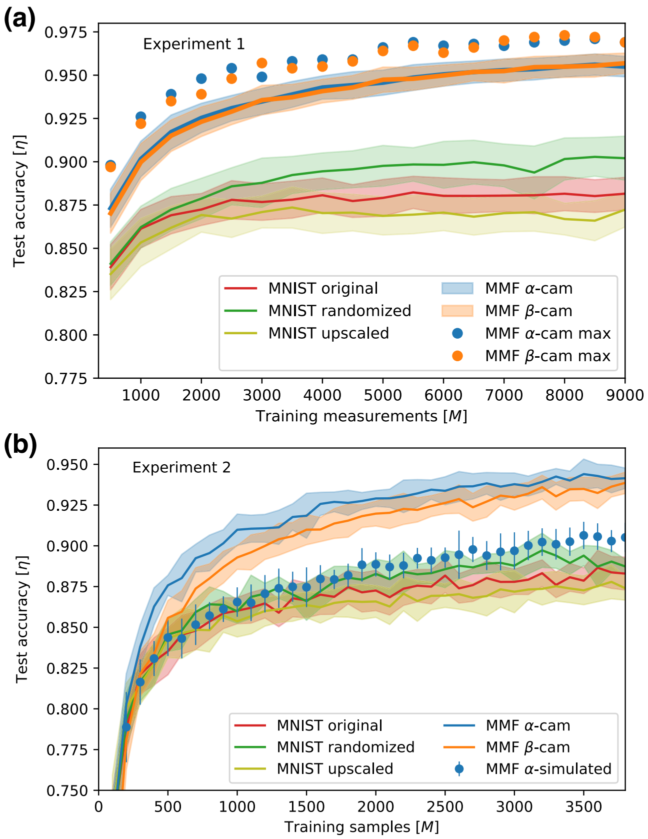

FIG. 2. Logistic regression performance using different training datasets. (a) Experiment 1, in which the model was trained with up to 9000 images. The original MNIST dataset ($l = 28$) was used as the reference performance (red line) for the logistic regression, together with its upscaled (light green) and randomized (green) versions. Training the classifier with MMF-transformed speckle images (blue and orange curves) resulted in the highest accuracy. These curves are obtained after 100 independent trainings, and with the dots we also report the corresponding maximum accuracy achieved per each fiber's facet. (b) Experiment 2, in which the model was trained using up to 3800 images. Compared to panel (a), we also include the output of the simulated fiber. The simulation was conducted by recovering the transmission matrix of the optical element and using a complex linear transmission model. The simulated model performed better than the MNIST dataset but did not reach the same performance as its experimental counterpart.

interactions of the MMFs, and the performance gain provided by the fine tuning of the hyperparameters with respect to each dataset would not change the main message of our work.

In experiment 2, we take a different static configuration of the fiber (i.e., characterized by another realization of $\mathbb{T}$), which we probe with an alternating sequence of random and MNIST images. Contrasting with experiment 1, here, we use the random patterns in the input (and the related projection) to characterize the transmission matrix of the fiber using the SmoothGS protocol [39]. We do this so that we can use the inferred $\mathbb{T}$ to simulate the propagation of the MNIST dataset through the fiber, obeying Eq. (2), and compare the classification performance of the linear model against that of the actual experimental measurements. To make a fair comparison with the simulated data, we tune the α-cam exposure time to avoid saturated measurements. Interestingly, we found that training the logistic regression with the α-simulated speckles does not perform well like the measured data. The accuracy achieved is better than the direct MNIST dataset but worse than that obtained using the experimental speckles [Fig. 2(b)]. We observe, then, a reality gap that may be due to the presence of noise and other experimental nonlinearities, which are not included in the way we model the physics of the system of Eq. (3) at low power. It may be conjectured that nonlinearities, which have also been studied in the framework of computational optics with much more intense pulsed light [26,43], also contribute at the lower intensities that we have been using in our experiments. Compared to the setup used in Ref. [26], we employed an energy density almost three orders of magnitude lower, also determined by the fact that we employed MMFs with large cores of 1 mm. On the other hand, the β-cam data was intentionally strongly underexposed (see Appendix E). By doing this, we found that considerable thresholding has only a marginally negative impact on the performance. Even when the camera loses most of its signal, the accuracy of the classifier drops by only ∼2% when compared with the better filling of the camera dynamic range in Fig. 2(b). This small performance drop enforces the idea that the MMF provides a class of random transformations that are particularly robust for carrying out classification tasks.

B. Accuracy of random projections and behavior of the training error

With this study, we have set a testing ground for different random projectors used to pretrain using the MNIST dataset, looking for those enhancing classification performance. To understand why the best-accuracy results are obtained with MMFs, we study a measure of the flatness of the energy landscape (i.e., the training error) around the different model solutions. Flatness is supposed to correlate well with generalization properties [7,44–48], meaning that it can provide insights into how the geometry of the projected space influences the classification errors of new data points. We use the method of the *local energy* to measure the flatness (see Ref. [7] and references therein), which involves adding multiplicative Gaussian noise to the model parameters, sampling configurations with a given noise, and eventually computing the average fraction of misclassified data points (see Appendix B for details). Performing this procedure to increase noise values yields an estimate of the flatness of the reference configuration. In Fig. 3, we can see that the local energy profile correlates

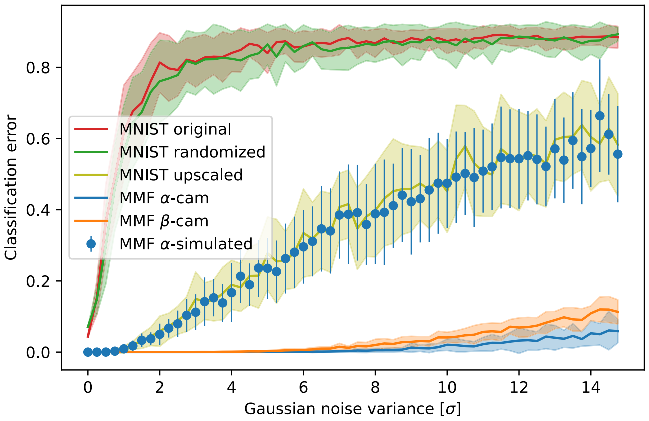

FIG. 3. Local energy profiles for models trained on the different projected datasets. Each point corresponds to the training error of configurations sampled with multiplicative noise around the reference, averaged over 30 samples. The reference configurations are models trained on 3800 examples. The error bars show the standard deviations of the error distribution.

well with test accuracy shown in Fig. 2: the flatter the solutions, the better the test accuracy. The only exception to this is the upscaled dataset, which has the same local energy profile as the simulated dataset but shows a lower test accuracy (we discuss this point in Sec. IV). A remarkable feature of the local energy profiles of MMF solutions is that they appear stable up to noise of the order of ten times the signal-to-noise ratio. This robustness to noise might be the reason for the excellent generalization performance on previously unseen data. This evidence supports the idea that MMFs are promising candidates for optical-neural-network computing. The models trained on MMF-projected data show very low local energy variation. On the one hand, this confirms the current idea in the literature that flatness correlates with generalization; on the other hand, it raises the question of why MMFs exhibit such a conceptual difference from their idealized model. This reality gap could signal the presence of something not yet taken into account in the theoretical description of the physics of experimental setups with MMFs used in low-power mode.

C. Real fiber propagation versus transmission-matrix simulation

The MMF is typically treated as a linear complex random projector, and its transmission rule can be estimated by finding the transmission matrix. In the case of a good $\mathbb{T}$ recovery, one would expect that the speckles simulated given a certain input will closely match the experimentally recorded output from the camera. Consequently, training a classifier with the simulated output should give a performance that is similar to that obtained with the real data. However, Fig. 2(b) highlights a strong discrepancy in accuracy with respect to the simulations, and Fig. 3 suggest a different local energy profile. This is a surprising fact that is worth investigating further. For this qualitative analysis, we use the data from experiment 2, which was specifically designed to recover the transmission matrix.

In Fig. 4(a), we show a representative output speckle pattern recorded by camera α. For better clarity, we restrict our analysis to a portion of the whole speckle output, identified with a red box and shown in Fig. 4(b). The result of the simulation is reported Fig. 4(c), which displays the reconstructed speckle pattern originating from the random input pattern that was included in the training set. Another representative pattern, not included in the training, is shown in Fig. 4(e), together with its corresponding simulated version in Fig. 4(f). For both, we observe minimal discrepancies between the real and simulated data, which we can quantify by plotting the difference maps between the two [Figs. 4(d) and 4(g)]. As an additional check, we also compute the focusing operator $\mathbb{T}\mathbb{T}^{\dagger}$, which we report in Fig. 4(h). The diagonality of the norm of this operator is normally used for testing the fidelity of the recovered transmission matrix [21]. In Fig. 5, we also compare the distribution of measured and simulated speckle intensities, computing the 2D histogram distribution [Fig. 5(a)] and its relative marginalizations [the histograms of the intensity distributions for each dataset, which are the integral of the 2D histogram along the two directions, Figs. 5(b) and 5(c)]. For completeness, since the 2D histogram is normally used to calculate the mutual information between the two datasets, we also report its value. Additionally, we analyzed the average autocorrelation of the speckles both from the measured data and the synthetic data created using the inferred transmission matrix [Figs. 5(d)–5(f)]. From the histogram analysis, a perfect match between the measured and simulated data would have produced a 2D histogram map with only diagonal entries. The fact that the diagonal is broadened implies that the correspondence between the measured intensities and the simulated dataset is not entirely captured by the recovery of the linear transmission, even if the speckles are effectively reproduced (as shown in Fig. 4).

To further restrict the reason for this discrepancy, we analyzed the average speckle autocorrelation of the measured [Fig. 5(d)] and simulated [Fig. 5(e)] datasets. We notice that the overall autocorrelation shape is very similar, and the profile plot in Fig. 5(f) confirms the close matching between the datasets. Since the autocorrelation is directly connected with the average size of the coherence region of a single speckle grain, having the same autocorrelation implies that the two speckle patterns have the same statistical spatial distribution, meaning they could accommodate a comparable number of optical modes. As an additional check, we decided to simulate the speckle output using a random-phase [flat distribution $\in [0 - 2\pi)$], complex-valued transmission matrix (keeping the modulus as retrieved in the experiments) and test its classification

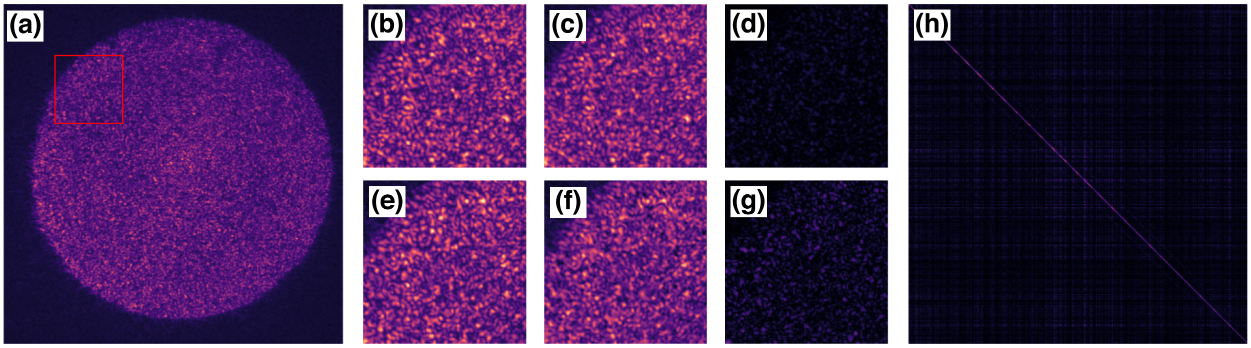

FIG. 4. (a) MMF α-cam speckle output after fiber propagation in experiment 2. The red box highlights a subregion magnified in panel (b) taken from the training dataset. (c) Simulated speckle output after transmission-matrix recovery, and (d) absolute difference between real and simulated speckle patterns. Using Eq. (E1), we can compare the average similarity of the measured and simulated speckles of all seen random modes, obtaining $\rho_{\text{train}} = 0.865 \pm 0.082$. (e) Speckle output recorded from the test set (not used for training), (f) corresponding simulated output using the recovered transmission, and (g) absolute difference between real and simulated data. Similarly, the average similarity of all unseen random modes is $\rho_{\text{test}} = 0.775 \pm 0.059$. (h) Focusing operator calculated using the recovered transmission of the output channels involved in the formation of the speckle in the red box.

performance. This new dataset performs similarly to the randomized MNIST (see supplementary code in the online repository), not reaching the experimental results.

D. Influence of the number of modes

As a last analysis, we evaluate the effect of the number of output modes in two different ways. In Fig. 6(a), we evaluate the effect of downscaling the MMF output of experiment 2 and, in Fig. 6(b), cropping it to a smaller window of increasing size. The effect of these operations is that we vary the size L of the output dataset used to train the classifier and, accordingly, the total number of output modes $N = L^2$. For both camera detections, reducing the number of modes has a negative impact on the performance, with the effect of the cropping operation being more drastic than that of rescaling. At around $L = 400$ pixels, however, both operations have similar effects, with performance nearly identical to the full-resolution image but with reduced numerical complexity. The fact that the output downscaled by a factor of around 2 has similar performance to the full-resolution dataset seems in agreement with the fact that the spatial correlation of the speckle pattern is wider than a single pixel in the detected image, thus introducing redundant information that can be compressed. We report, however, that this also happens with the cropped version of the output, which still shares the same spatial properties of the average speckle size. Remarkably, we also register that the fiber simulation does not perform equally well, with the only exception being at very small sizes (up to $L = 36$), when the accuracy is still low and of no practical use. Furthermore, we notice that the other datasets (randomized and upscaled) still perform worse compared to the hardware fiber after $L = 54$, even though the accuracy obtained in this regime is relatively low. Additionally, from Fig. 6(a), we observe that upscaling the original MNIST data has a negative impact on the performance, possibly due to overfitting, as the ideal dimension of the dataset sits at around $L = 18$–32 (local maximum of the curve). This also explains the lower performance registered in Fig. 2. On the other hand, in Fig. 6(b), the same dataset has a dramatic dependence on cropping. This is to be expected because by cropping we are restricting the observation window down to a small feature of the number image and not capturing its entire shape. Among these options, we can operatively conclude that best way to improve classification accuracy is by using a hardware MMF projector.

IV. PERSPECTIVES

In this work, we used MMFs to realize random transformations of the MNIST dataset showing that a linear classifier has better accuracy on the MMF-transformed dataset than on the original one. Complementary to high-intensity pulsed excitation [26], this transformation (MMF and camera detection) is nonlinear, even in the continuous low-power regime, and it increases the dimension of the data, but those characteristics alone are not enough to explain the improved accuracy. In fact, data upscaling (which increases the dimension), random matrix multiplication (which projects on random space), and the MMF simulation did not reach performances similar to the transformation provided by the physical MMF. As noted in Sec. III, our goal was not to compete with the accuracy of more sophisticated architectures, but rather to show that MMFs are simple—yet robust—hardware solutions for optical computing. For example, convolutional neural networks exploit spatial correlations in the data and work particularly well for image datasets. Instead, our

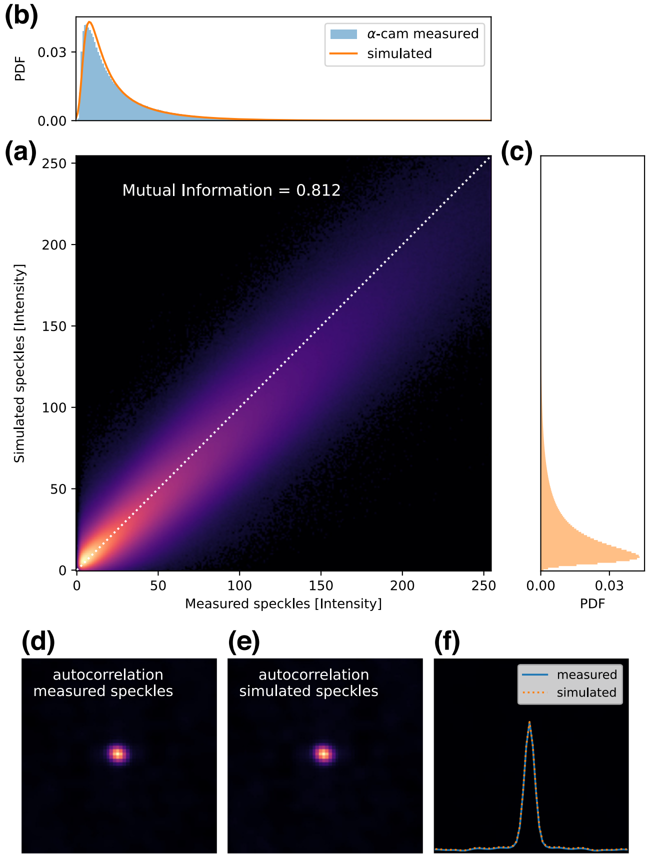

FIG. 5. (a) Bidimensional intensity histogram between measured and simulated speckles. The diagonal is the ideal histogram map when the simulation perfectly matches the measured data. Instead, one can notice that dispersion occurs, quantified by a mutual information value of 0.812, cf. Eq. (E2). (b) Histogram plot of the measured speckle intensities, projection along the vertical axis of the 2D histogram. In orange, we superimpose the plot of the histogram of the simulated speckles. (c) Intensity histogram of the simulated speckles, 2D histogram projection along horizontal axis. (d) Average autocorrelation of the measured speckles, and (e) autocorrelation of the simulated speckle pattern using the inferred $\mathbb{T}$. (f) Autocorrelation difference (dark image in the background) and central-profile plot of the two functions, demonstrating practically identical average speckle sizes recovered after the transmission characterization.

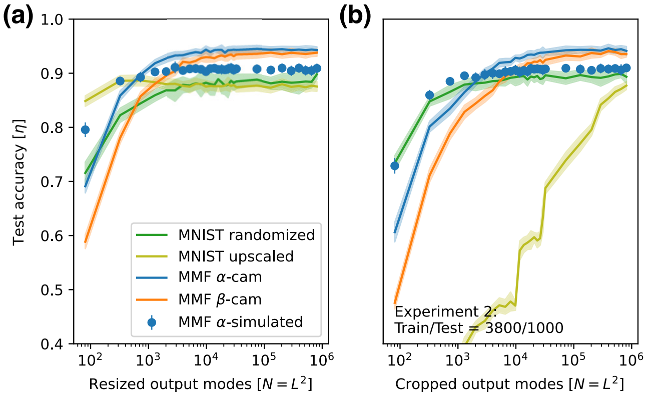

FIG. 6. Training accuracy trade-off when reducing the number of output fiber modes in experiment 2. We study the performances obtained using demagnified camera measurements as a function of their output size (blue and orange plots). With the dots, we report the same study performed with simulated speckle patterns. The bright and olive green, respectively, show the results for the upscaled and randomized MNIST datasets. In (a), the study is done by resizing the output patterns, and in (b), a similar study is done by cropping windows smaller than the original dimension down to different sizes L (thus excluding peripherical speckles). Along the x axis, we report the maximum number of optical modes allowed after resizing and cropping, $N = L^2$. For both, we notice that performances remain stable down to a substantial reduction of the number of modes used in the training set (around 400 pixels, 80% fewer pixels than the full-resolution dataset).

approach is closer to that of a fixed-weight densely connected network, leaving room for applicability to a variety of different data types; however, in contrast to general random transformations (which destroy spatial correlations), the fiber output presents a correlation property determined by the average size of the speckle patterns.

As a physical neural network, an MMF is cheap, can be flexibly mounted to deliver light to a user-defined position, and offers a different set of random projections each time it is repositioned (thus requiring independent training of the output layer). Indeed, we still need an SLM and at least one camera to record the speckled projection, but these are almost unavoidable in any ONN configuration. To the best of our knowledge, the current state of the art is achieved using field-programmable gate array hardware in conjunction with data augmentation, reaching 98% test accuracy [49]. Another approach using disordered optical media exploits polaritons to reach 96% accuracy [50], which is comparable with average optimizations obtained using the MMF approach. These results strengthen the notion that MMFs are promising tools for neuromorphic computing, with the additional advantage of their simplicity and ease of use. Further, we believe that our results could be relevant for the theoretical understanding of deep neural networks: in the spirit of random-feature models [5–7], we showed that the class in which we sample the random features plays important role in the accuracy, as suggested in Ref. [8]. In fact, while taking a Gaussian random matrix already improves the accuracy somewhat, the transformation implemented by an MMF makes a much bigger difference. Further investigation is needed to understand why the specific hardware transformation provided by the MMF is so effective. In particular, the local energy profiles suggest that this effectiveness could be explained by studying the wide flat regions in the loss landscape, in the same spirit as in Ref. [7]: to do so, the authors use a quantity called *local entropy*, which is only approximated

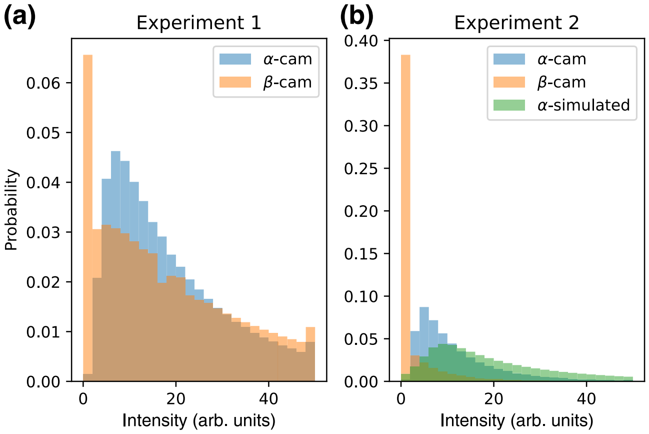

FIG. 7. Intensity distributions of the speckle patterns measured in cameras α and β during two independent experiments. Given a stable laser output, we modify the camera exposure time to force a certain amount of nonlinearity (in the form of a recording threshold) in the measurement process. (a) In experiment 1, the dynamic range of camera α fits the intensity distribution of the speckle images recorded well, whereas camera β underexposes around 7% of the signal. (b) In experiment 2, camera α has a similar trend to that of experiment 1, but in camera β, we strongly underexpose the images, cutting out 37% of the light intensity reaching the sensor.

by the local energy that we discussed here (this might explain the discrepancy between the local energy profile of the upscaled dataset in Fig. 3 and its test error in Fig. 2).

Here, we put forward some conjectures based on the present study. First, the fact that the accuracy gap between the physical MMF data and its simulation (Fig. 2) is reflected in the local energy profile (Fig. 3) makes us confident that the two approaches indeed belong to different classes of random transformation. The fact the physical MMF transformation is so robust to perturbations is consistent with the great redundancy of the data that emerges from Figs. 6 and 7, in which we see that we can delete the majority of the signal before losing accuracy. We conjecture that the random transformation realized by MMFs leads to well-separated projections in the high-dimensional space, which allow for good classification accuracy that is also resistant to noise in a way that is reminiscent of error-correcting codes. All these considerations highlight the need to further investigate how MMF devices can be modeled and exploited, particularly in the design of optical neural networks.

The code to reproduce the results in Fig. 2 is freely downloadable from GitHub at Ref. [51], and the relative datasets are available from FigShare [52].

ACKNOWLEDGMENTS

We thank Dr. Raffaele Marino for useful discussions. We acknowledge the support of LazioInnova - Regione Lazio under the program *Gruppi di ricerca 2020* - POR FESR Lazio 2014–2020, Project NanoProbe (Application code A0375-2020-36761), the support from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program, Project LoTGlasSy (Grant No. 694925, Prof. Giorgio Parisi), and the support of the PRIN project "Complexity, disorder and fluctuations" of the Italian Ministry of University and Research (MUR) under the NextGenerationEU program, Grant No. 2022LMHTET.

APPENDIX A: NEURAL-NETWORK ARCHITECTURE AND TRAINING PROCEDURE

Our classification model involves a potentially fully connected layer (in the sense that we do not restrict mode couplings of any intensity) that linearly maps the 28 × 28 image space into a higher-dimensional $N = 900 \times 900$ output space. On the output space, we build a linear classification model using the `LogisticRegression` function provided by the Python library RAPIDS AI [53], the GPU equivalent of the scikit-learn implementation. Given an input pattern $\{\xi_i\}_{i=1,\ldots,N}$ the `LogisticRegression` function performs a weighted average of the N input channels, producing a score $z_j = \sum_{i=1}^{N} w_{ji}\xi_i$ for each of the ten classes corresponding to each type of digit. The scores are then transformed to probabilities with

$$p_j = \frac{e^{z_j}}{\sum_{j'=1}^{10} e^{z_{j'}}} \qquad \text{(A1)}$$

and plugged into a cross-entropy loss function that is a sum of the contributions coming from all of the P input patterns that we are using to train the model:

$$L(\mathbf{w}) = -\sum_{\mu=1}^{P} \log(p_{j^*}), \qquad \text{(A2)}$$

where j^* is the index of the correct class of each input pattern. The cross-entropy loss $L(\mathbf{w})$ is then minimized with a gradient-descent-related strategy to find the configuration of the weights $\mathbf{w}^*$ that has the highest classification accuracy. The classification accuracy is defined as the fraction of correctly classified entries divided by the total number of training (or test) images. To compute this after the parameter optimization, we make use of the `sklearn.metrics.accuracy_score` function of the scikit-learn library.

APPENDIX B: MEASURE OF LOCAL ENERGY

Given a loss (namely energy) function L that depends on a set of parameters $\mathbf{w}$, we define the local energy as the expectation value

$$L_{\texttt{local}}(\sigma) = \mathbb{E}_{\eta_{ij}\sim\mathcal{N}(0,\sigma)} L(\{w_{ij}\,\eta_{ij}\}), \qquad \text{(B1)}$$

where $\{\eta_{ij}\}$ is a set of independent and identically distributed random Gaussian variables with zero mean and variance σ that multiply element-wise the set model parameters $\{w_{ij}\}$. The local energy $L_{\texttt{local}}(\sigma)$ still depends on the variance σ of the Gaussian noise. As explained in the main text, we are interested in studying how quickly the local energy increases when we increase σ: from the literature (see main text), we know that a slower increase is correlated with a higher test accuracy. For Fig. 3 of the main text, we choose L as the fraction of misclassified data points in the training set.

APPENDIX C: EXPERIMENTAL SETUP

A sketch of the experimental setup is shown in Fig. 1(c). In the experiments, we use a continuous Melles Griot He-Ne laser (632.8 nm) as the light source. The emitted beam is magnified 15 times through a 5 : 75-cm telescope before being imprinted on a Hamamatsu SLM in polarization configuration (model LCOS-SLM x10488 series, pixel size: 20 μm). The real-space plane of the SLM is then recreated on the entrance facet of the optical fiber using a pair of 50 : 7.5-cm focal lenses after the spatially modulated beam profile has been collected. A Thorlabs FT1000EMT, NA = 0.39, 1-m long, 1-mm core multimode optical fiber is used. We indicate the facets of the MMF with the letters α and β. Two IDS cameras (UI-5370CP-M-GL and UI-5480CP-M-GL) with pixel sizes of 5.5 and 2.2 μm, respectively, are used to collect the counter-polarized (with respect to the laser) reflection from the injection surface as well as the transmission signal. To achieve the same spatial resolution of 1.1 μm /pixel on both cameras, the magnifications are set to 5× and 3×, respectively. The MNIST handwritten digits and random masks are sent to the SLM in alternated sequences and are encoded in the same way. In practice, for each of these, we send an image (random or MNIST) having a size of 28 × 28 pixels, focusing it so that it is inscribed on the input facet of the optical fiber. Each pixel uses grayscale values ranging from 0 to 10. The random patterns are sent for the sole purpose of characterizing the fiber transmission, and they are not used for training of the classification layer. The light propagating through this disordered optical device reaches both edges and produces a seemingly random interference pattern of intensities (the speckles).

APPENDIX D: NUMBER OF OPTICAL MODES

The fiber used (FT1000EMT, Thorlabs) has a diameter of $d = 1$ mm with NA = 0.39. Thus, the maximum theoretical number of supported modes is $N_{\text{modes}} = (\pi d\text{NA}/\lambda)^2/2$, which gives around 1.871×10^3 modes. For the experimental realization, the number of optical modes is influenced by the number of camera pixels used to record the fiber's output and the average physical size of the speckles. In our case, the average full-width half maximum of the speckles is 1 pixel, and using a squared portion of the central core of the fiber having $L = 600$ determines a maximum total number of imaged modes equal to $L^2 = 360 \times 10^3$ modes. This is a reduced fraction of the total number of imaged modes of the entire facet, consisting of about 635×10^3 modes.

APPENDIX E: UNDEREXPOSURE, CAMERA SATURATION, AND MEASUREMENT STABILITY

When setting the exposure time of the camera, we are implicitly acting on the way it records the signal. If the exposure time is fast enough with respect to the intensity delivered, the camera underexposes the signal, i.e., it does not detect the signal in a particular region. The opposite effect, overexposure, happens when the intensity is too high for a long exposure of the image. In both cases, a nonlinear threshold is introduced in the detected signals. To try to assess the effect of this on the classifier accuracy, we tried to explore several intensity distributions of the datasets recorded in the camera.

In experiment 1, Fig. 7(a), α-cam provides an optimal dynamic range, with both low underexposure (0.1%) and overexposure (1%). Instead, β-cam recorded the signal underexposing 7% of the total pixels in the image. In experiment 2, Fig. 7(b), α-cam correctly samples the intensities, whereas β-cam is set to cut off 37% of the pixels. Additionally, we report the intensity distribution obtained with the simulation of the light propagating through the fiber and detected by α-cam. We notice a substantial difference between the intensity distributions of the recorded and simulated data: this could explain the different performances achieved by the two datasets.

Over the entire duration of the experiment, we continuously monitor the fiber stability by sending an identical image to the SLM. When the fiber is sufficiently stable, the speckle patterns produced at the facets must always be identical to those recorded at the beginning of the experiment. Keeping the camera frame $\tau = 0$ as a reference for both cameras, we compute the normalized scalar product against the speckle image at a given time τ':

$$\rho\left(\tau, \tau'\right) = \frac{s_{\tau}^{\{\alpha,\beta\}} \cdot s_{\tau'}^{\{\alpha,\beta\}}}{\left|s_{\tau}^{\{\alpha,\beta\}}\right|\left|s_{\tau'}^{\{\alpha,\beta\}}\right|}, \qquad \text{(E1)}$$

where s is the recorded speckle pattern at each time. Using this metric, $\rho \approx 1$ means the measurements are highly correlated, whereas $\rho \approx 0$ implies that the system is decorrelated during the measurement. Figure 8 shows a stability study across the entire duration of experiment 1. We notice that α-cam remains highly correlated (96% minimum)

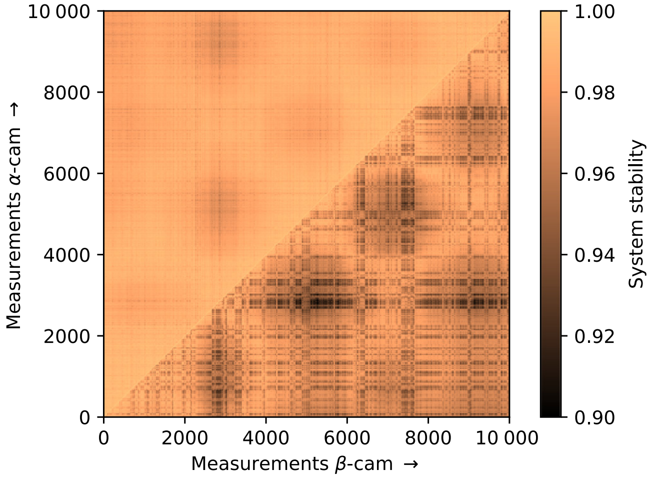

FIG. 8. Measurement stability during experiment 1. Keeping the initial probing frame, we compute the normalized scalar product against the output of the same frame at different times for both cameras. The upper half of the plot is the α-cam correlation stability (hence how similar the output looks when the same input is sent again during the experiment) and the bottom half shows that for β-cam. In both cases, the correlation is higher than 90%.

compared to β-cam (90% minimum). Despite the lower correlation stability and 7% underexposed pixel values, the β-cam results are as accurate as the α-cam results during the classification of the test set (Fig. 2).

In Fig. 5 we compare the output speckles corresponding to the same input through the real MMF and through a synthetic MMF whose transmission matrix is that inferred from the data by phase retrieval. The two do not appear to be the same, that is, their scatter plot is not exactly diagonal. We quantify their mutual difference by means of the mutual information

$$I(\text{real}|\text{synth}) \equiv \sum_{i=1}^{256} P_{\text{real}}(y_i) \log \frac{P_{\text{real}}(y_i)}{P_{\text{synth}}(y_i)}, \tag{E2}$$

where 256 is the number of intensity bins. A perfect match would yield $I = 1$, whereas in Fig. 5(a) we find $I = 0.812$.

[1] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical Learning with Sparsity, The Lasso and Generalizations* (CRC Press, Taylor & Francis Group, New York, NY, USA, 2015).

[2] David J. C. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[3] Lenaic Chizat, Edouard Oyallon, and Francis Bach, On lazy training in differentiable programming, Adv. Neural Inf. Process. Syst. **32**, 2937 (2019).

[4] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart, Disentangling feature and lazy training in deep neural networks, J. Stat. Mech.: Theory Exp. **2020**, 113301 (2020).

[5] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová, in *International Conference on Machine Learning* (PMLR, 2020), p. 3452.

[6] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová, in *Mathematical and Scientific Machine Learning* (PMLR, 2022), p. 426.

[7] Carlo Baldassi, Clarissa Lauditi, Enrico M. Malatesta, Rosalba Pacelli, Gabriele Perugini, and Riccardo Zecchina, Learning through atypical phase transitions in overparameterized neural networks, Phys. Rev. E **106**, 014116 (2022).

[8] Gabriele Perugini, *et al.*, (manuscript in preparation).

[9] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A. B. Miller, and Demetri Psaltis, Inference in artificial intelligence with deep optics and photonics, Nature **588**, 39 (2020).

[10] Lorenzo De Marinis, Marco Cococcioni, Piero Castoldi, and Nicola Andriolli, Photonic neural networks: A survey, IEEE Access **7**, 175827 (2019).

[11] Xing Lin, Yair Rivenson, Nezih T. Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan, All-optical machine learning using diffractive deep neural networks, Science **361**, 1004 (2018).

[12] Tiankuang Zhou, Xing Lin, Jiamin Wu, Yitong Chen, Hao Xie, Yipeng Li, Jingtao Fan, Huaqiang Wu, and Lu Fang, and Qionghai Dai, Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit, Nat. Photonics **15**, 367 (2021).

[13] Jingxi Li, Deniz Mengu, Nezih T. Yardimci, Yi Luo, Xurong Li, Muhammed Veli, Yair Rivenson, Mona Jarrahi, and Aydogan Ozcan, Spectrally encoded single-pixel machine vision using diffractive networks, Sci. Adv. **7**, eabd7690 (2021).

[14] Yaser S. Abu-Mostafa and Demetri Psaltis, Optical neural computers, Sci. Am. **256**, 88 (1987).

[15] Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, *et al.*, Deep learning with coherent nanophotonic circuits, Nat. Photonics **11**, 441 (2017).

[16] Jonathan Dong, Mushegh Rafayelyan, Florent Krzakala, and Sylvain Gigan, Optical reservoir computing using multiple light scattering for chaotic systems prediction, IEEE. J. Sel. Top. Quantum Electron. **26**, 1 (2019).

[17] Yi Luo, Deniz Mengu, Nezih T. Yardimci, Yair Rivenson, Muhammed Veli, Mona Jarrahi, and Aydogan Ozcan, Design of task-specific optical systems using broadband diffractive neural networks, Light Sci. Appl. **8**, 112 (2019).

[18] James Spall, Xianxin Guo, and A. I. Lvovsky, Hybrid training of optical neural networks, Optica **9**, 803 (2022).

[19] Ying Zuo, Bohan Li, Yujun Zhao, Yue Jiang, You-Chiuan Chen, Peng Chen, Gyu-Boong Jo, Junwei Liu, and Shengwang Du, All-optical neural network with nonlinear activation functions, Optica **6**, 1132 (2019).

[20] Logan G. Wright, Tatsuhiro Onodera, Martin M. Stein, Tianyu Wang, Darren T. Schachter, Zoey Hu, and Peter

L. McMahon, Deep physical neural networks trained with backpropagation, Nature **601**, 549 (2022).

[21] Sebastien M. Popoff, Geoffroy Lerosey, Rémi Carminati, Mathias Fink, Albert Claude Boccara, and Sylvain 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).

[22] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose, Recent advances in physical reservoir computing: A review, Neural Netw. **115**, 100 (2019).

[23] Uttam Paudel, Marta Luengo-Kovac, Jacob Pilawa, T. Justin Shaw, and George C. Valley, Classification of time-domain waveforms using a speckle-based optical reservoir computer, Opt. Express **28**, 1225 (2020).

[24] Xavier Porte, Anas Skalli, Nasibeh Haghighi, Stephan Reitzenstein, James A. Lott, and Daniel Brunner, A complete, parallel and autonomous photonic neural network in a semiconductor multimode laser, J. Phys.: Photonics **3**, 024017 (2021).

[25] Ryosuke Takagi, Ryoichi Horisaki, and Jun Tanida, Object recognition through a multi-mode fiber, Opt. Rev. **24**, 117 (2017).

[26] Uğur Teğin, Mustafa Yıldırım, Christophe Moser, and Demetri Psaltis, Scalable optical learning operator, Nat. Comput. Sci. **1**, 542 (2021).

[27] Alaa Saade, Francesco Caltagirone, Igor Carron, Laurent Daudet, Angélique Drémeau, Sylvain Gigan, and Florent Krzakala, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2016), p. 6215.

[28] Chao Du, Fuxi Cai, Mohammed A. Zidan, Wen Ma, and Seung Hwan Lee, and Wei D. Lu, Reservoir computing using dynamic memristors for temporal information processing, Nat. Commun. **8**, 1 (2017).

[29] Dario Ballarini, Antonio Gianfrate, Riccardo Panico, Andrzej Opala, Sanjib Ghosh, Lorenzo Dominici, Vincenzo Ardizzone, Milena De Giorgi, Giovanni Lerario, Giuseppe Gigli, *et al.*, Polaritonic neuromorphic computing outperforms linear classifiers, Nano Lett. **20**, 3506 (2020).

[30] Tao Chen, Jeroen van Gelder, Bram van de Ven, Sergey V. Amitonov, Bram De Wilde, Hans-Christian Ruiz Euler, Hajo Broersma, Peter A. Bobbert, Floris A. Zwanenburg, and Wilfred G. van der Wiel, Classification with a disordered dopant-atom network in silicon, Nature **577**, 341 (2020).

[31] Daniele Ancora and Luca Leuzzi, Transmission matrix inference via pseudolikelihood decimation, J. Phys. A: Math. Theor. **55**, 395002 (2022).

[32] Elbert G. van Putten and Allard P. Mosk, The information age in optics: Measuring the transmission matrix, Physics **3**, 22 (2010).

[33] Sébastien Popoff, Geoffroy Lerosey, Mathias Fink, Albert Claude Boccara, and Sylvain Gigan, Image transmission through an opaque material, Nat. Commun. **1**, 81 (2010).

[34] Jochen Aulbach, Bergin Gjonaj, Patrick M. Johnson, Allard P. Mosk, and Ad Lagendijk, Control of light transmission through opaque scattering media in space and time, Phys. Rev. Lett. **106**, 103901 (2011).

[35] S. M. Popoff, A. Goetschy, S. F. Liew, A. D. Stone, and H. Cao, Coherent control of total transmission of light through disordered media, Phys. Rev. Lett. **112**, 133903 (2014).

[36] Stefan Rotter and Sylvain Gigan, Light fields in complex media: Mesoscopic scattering meets wave control, Rev. Mod. Phys. **89**, 015005 (2017).

[37] Angélique Drémeau, Antoine Liutkus, David Martina, Ori Katz, Christophe Schülke, Florent Krzakala, Sylvain Gigan, and Laurent Daudet, Reference-less measurement of the transmission matrix of a highly scattering material using a DMD and phase retrieval techniques, Opt. Express **23**, 11898 (2015).

[38] Guoqiang Huang, Daixuan Wu, Jiawei Luo, Liang Lu, Fan Li, Yuecheng Shen, and Zhaohui Li, Generalizing the Gerchberg–Saxton algorithm for retrieving complex optical transmission matrices, Photonics Res. **9**, 34 (2021).

[39] Daniele Ancora, Lorenzo Dominici, Antonio Gianfrate, Paolo Cazzato, Milena De Giorgi, Dario Ballarini, Daniele Sanvitto, and Luca Leuzzi, Speckle spatial correlations aiding optical transmission matrix retrieval: The smoothed Gerchberg–Saxton single-iteration algorithm, Photonics Res. **10**, 2349 (2022).

[40] The underlined network is one layer deep because it can be described by a single matrix multiplication, as in Eq. (1).

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE **86**, 2278 (1998).

[42] Sanghyeon An, Minjun Lee, Sanglee Park, Heerin Yang, and Jungmin So, An ensemble of simple convolutional neural network models for MNIST digit recognition, arXiv:2008.10400.

[43] Ilker Oguz, Jih-Liang Hsieh, Niyazi Ulas Dinc, Uğur Teğin, Mustafa Yildirim, Carlo Gigli, Christophe Moser, and Demetri Psaltis, Programming nonlinear propagation for efficient optical learning machines, Adv. Photonics **6**, 016002 (2024).

[44] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio, Fantastic generalization measures and where to find them, arXiv:1912.02178.

[45] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina, Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[46] Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. **113**, E7655 (2016).

[47] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina, Entropy-SGD: Biasing gradient descent into wide valleys, J. Stat. Mech.: Theory Exp. **2019**, 124018 (2019).

[48] Fabrizio Pittorino, Carlo Lucibello, Christoph Feinauer, Gabriele Perugini, Carlo Baldassi, Elizaveta Demyanenko, and Riccardo Zecchina, Entropic gradient descent

algorithms and wide flat minima, J. Stat. Mech.: Theory Exp. **2021**, 124015 (2021).

[49] Alejandro Morán, Christiam F. Frasser, Miquel Roca, and Josep L. Rosselló, Energy-efficient pattern recognition hardware with elementary cellular automata, IEEE Trans. Comput. **69**, 392 (2020).

[50] Rafał Mirek, Andrzej Opala, Paolo Comaron, Magdalena Furman, Mateusz Król, Krzysztof Tyszka, Bartłomiej Seredynski, Dario Ballarini, Daniele Sanvitto, and Timothy C. H. Liew, Neuromorphic binarized polariton networks, Nano Lett. **21**, 3715 (2021).

[51] https://github.com/danieleancora/MMFclassification.git

[52] https://doi.org/10.6084/m9.figshare.25551186.v1

[53] RAPIDS Development Team, RAPIDS: Collection of Libraries for End to End GPU Data Science, https://rapids.ai (2018).