EPFL

Thèse n° 9406

Learning of physical systems: from inference to control

Présentée le 31 mars 2022

Faculté des sciences et techniques de l'ingénieur
Laboratoire de dispositifs photoniques appliqués
Programme doctoral en génie électrique

pour l'obtention du grade de Docteur ès Sciences

par

Babak RAHMANI

Acceptée sur proposition du jury

Prof. M. Unser, président du jury
Prof. C. Moser, directeur de thèse
Prof. A. Ozcan, rapporteur
Dr L. Su, rapporteur
Prof. M. Liebling, rapporteur

École
polytechnique
fédérale
de Lausanne

2022

To my family...

Acknowledgements

I would like to take this opportunity to extend my gratitude to all the colleagues and friends who contributed, in one way or another, to the completion of my Ph.D. and realization of this thesis.

First and foremost, thanks to Professor Christophe Moser. I am very grateful for having had him as my thesis director. I really enjoyed working, discussing and brainstorming with him. I thank him for all the advice, both personal and carrier-wise, he gave me over these past four years. Special thanks also go to Prof. Psaltis who was like a co-advisor to me. I enjoyed all the meetings and discussions we had together.

I want to thank my office mates, Chiara, Maya, Enrico and Jorge for their friendship, fun discussions, and many more. A thanks also goes to every other colleagues in the lab: Jan, Antoine, Leo, Damien, Paul, Tim, Mathieu, Eirini, Giulia and Ilker.

Last but not least, I would like to thank my family back in Iran who supported me. Also, I thank my friends in Lausanne and overseas: Farzad Pourkamali, Salar Rahimi, Mohammadreza Ebrahimi, Massod Valavi, Amirhossein Saba, Alireza Mohammadshahi, Farnaz Eslamishoar, Shayan Khalooei, Yasmin Ghadyani and many others who made Ph.D. life more enjoyable.

Lausanne, March 2022 B. R.

Abstract

To characterize a physical system to behave as desired, either its underlying governing rules must be known a priori or the system itself be accurately measured. The complexity of full measurements of the system scales with its size. When exposed to real-world conditions, such as perturbations or time-varying settings, the system calibrated for a fixed working condition might require non-trivial re-calibration, a process that could be prohibitively expensive, inefficient and impractical for real-world use cases.

In this thesis, a learning procedure for solving highly ill-posed problems of modeling a system's forward and backward response functions is proposed. In particular, deep neural networks are used to infer the input of a system from partial measurements of its outputs or to obtain a desired target output from a physical system.

I showcase the applicability of the proposed methods for inference and control in optical multimode fibers. Amplitude/phase-encoded input of a multimode fiber is reconstructed from intensity-only measurements of the outputs. Conversely, the required input of the fiber for projecting a desired output is obtained using intensity-only measurements of the output.

Next, the stochastic neural network of the retina in Salamander is modeled by a probabilistic neural network. The model is used to optimize the input stimuli so as to find the simplest spatiotemporal patterns that elicit the same neuronal spike responses as those elicited by high-dimensional stimuli.

As demonstrated in this thesis, application of data-driven methods for characterization of complex large-scale real-world systems has proved useful in simplifying the measurement apparatus, end-to-end optimization of the system and automatic compensation of perturbation.

Keywords: machine learning, deep learning, imaging, forward modeling, multimode fibers, retina modeling, information bottleneck, system characterization.

Résumé

Pour caractériser un système physique pour qu'il se comporte comme on le souhaite, soit ses règles de gouvernance sous-jacentes doit être connu a priori ou le système lui-même doit être mesuré avec précision. La complexité du plein les mesures du système évoluent avec sa taille. Lorsqu'ils sont exposés à des conditions réelles, de tels comme des perturbations ou des paramètres variables dans le temps, le système est calibré pour une condition de travail fixe pourrait nécessiter un réétalonnage non trivial, un processus qui pourrait être d'un coût prohibitif, inefficace efficace et peu pratique pour les cas d'utilisation réels. Dans cette thèse, une procédure d'apprentissage pour résoudre des problèmes très mal posés de modélisation de la des fonctions de réponse avant et arrière sont proposées. En particulier, les réseaux de neurones profonds sont utilisés pour déduire l'entrée d'un système à partir de mesures partielles de ses sorties ou pour obtenir une sortie cible souhaitée d'un système physique. Je présente l'applicabilité des méthodes proposées pour l'inférence et le contrôle en optique fibres multimodes. L'entrée codée en amplitude/phase d'une fibre multimode est reconstruite à partir de mesures d'intensité uniquement des sorties. Inversement, l'apport requis de la fibre pour projeter une sortie souhaitée est obtenue en utilisant uniquement des mesures d'intensité de la sortie. Ensuite, le réseau de neurones stochastiques de la rétine de Salamander est modélisé par un modèle probabiliste réseau neuronal. Le modèle est utilisé pour optimiser les stimuli d'entrée afin de trouver le plus simple modèles spatio-temporels qui suscitent les mêmes réponses neuronales que celles suscitées par stimuli de haute dimension. Comme démontré dans cette thèse, l'application de méthodes basées sur les données pour la caractérisation des systèmes complexes à grande échelle du monde réel s'est avéré utile pour simplifier l'approche de mesure paratus, optimisation de bout en bout du système et compensation automatique des perturbations.

Mots clefs : machine learning, deep learning, imagerie, modélisation directe, fibres multimodes, modélisation de la rétine, goulot d'étranglement informationnel, caractérisation du système.

Contents

List of Figures

List of Tables

1 Introduction

How does one go about solving a problem in natural sciences? This process of problem solving usually starts with identifying some general rules governing the system in which the problem is defined. Next step is to understand the inputs to that system. What are they? How does the system respond to each input? The objective of the problem might be to obtain the input of the system given its corresponding output or to obtain the required input for producing a desired output. If the underlying rules governing the system are known a priori, depending on the complexity of the problem, it might be possible to theoretically provide answers to the above questions. To make it clear, let us illustrate with an example. In optical physics, one of the most basic problems is the study of light propagation in a particular medium. In order to predict the amplitude and phase of the light field in any plane, one can solve Maxwell equations or use approximations such as the Fresnel-Kirchoff diffraction formula. When the media features random scatterers that perturb the propagation of light at the scale of the wavelength, the problem is computationally complex and requires to solve the full vectorial wave equation. A less computationally intensive alternative is to consider the light propagation as linear, meaning that the polarizability of the medium is proportional to the amplitude of the light, which is often the case, and treat the light propagation as a linear system. In such a system, if one knows the amplitude and phase in two planes, described by the vector X and Y respectively, the two vectors can be easily related by a linear transform with a matrix T, $Y = TX$. However, providing experimental measurements for X and Y can be challenging. The characterization of the complex set of amplitude and phase values X and Y is possible with holographic phase measurements. However, the apparatus needed for such measurements is nontrivial as it requires to handle and correct for phase drifts of the interference pattern. Using amplitude only measurement via detecting the light field with a camera is much simpler experimentally, but the measurement does not produce the light amplitude and phase separately, thus the matrix T cannot be inferred . Thus, holographic phase measurement has been the gold standard method for characterizing light propagation in a scattering medium. The reason for that roots from the fact that with holographic measurements, the system is linear whereas in the latter case (intensity measurement), the system is nonlinear. Characterizing a linear system is obviously

simpler than a nonlinear one. So far, we saw that if the underlying rules governing the system is unknown or the system is too complex to be analytically modelled or when experimental realization of the problem requires a nontrivial measurement apparatus, resorting to other techniques might prove more advantageous. The latter is also necessary when the system is known but perturbations present in the experimental realization of the problem drastically modify the system.

In this thesis, I propose, for the first time, a statistical and data-driven method to describe the light propagation in multimode fibers whose light field output resembles the output of a scattering medium and where only the light intensity is measured. This system represents a well-defined non-linear system, whose complexity can be modified by suitable choice of the number of modes in the fiber. In this thesis, I present those data driven methods and provide a general framework that is applicable not only to optical physics but also can be generalized to other disciplines such as in neuroscience where I show its potential to characterize the vision system. Broadly speaking, data-driven methods, henceforth referred to as learning methods, are techniques that use observations from the system, for example samples of the inputs and outputs, to hypothesize a statement about the system that is corroborated by the data. Within this framework, the general approach for learning a physical system starts with (1) acquiring many samples from the system, for example obtaining input-output examples from the system. (2) choosing the sample size large enough to represent a general feature of the system. (3) proposing a quantifiable metric for learning (the error function) that depends on the application and the question that the experimentalist is trying to provide an answer for (4) choosing the learning method: parametric or non-parametric. For example, in parametric methods, usually the parameters of a chosen function f are learned to represent the data. Once these parameters are available, new data that were not used in the first place for obtaining the parameters are used to assess the generalizability of the function f on the new data.

Deep learning is a parametric method that is used for learning a general relation between sets of input-output data, by fitting a large number of learnable variables, known as weights and biases and referred to collectively as a neural network. Today, a plethora of neural network architectures ranging from the fully-connected network to more advanced Convolutional neural networks (CNNs) have been proposed. CNNs are a subclass of neural networks, which have been proposed and have shown better performance over other neural networks by decreasing the computation cost of fully connected layers through parameter sharing and use of sparse filters while, at the same time, increasing the number of layers in the network to achieve deep networks for solving more complex problems while speeding up computations. With this new computational power, CNNs have been used in various fields and applications for example, in optical microscopy to solve for phase recovery in non-linear inverse problems [1] [2] [3] and many other.

Problems which involve a system that accepts an input and produce a corresponding output in response to that input can be categorized into two main classes: inference and control problems. In the former, the goal is to infer the input of the system given several examples of

the outputs. In the latter, the goal is to obtain the appropriate input that produce a desired target output. In this case, I assume that the system could be probed as many times as needed. However, the experimentalist might have only access to partial measurements from the system. For example, in the optical physics problem, the experimentalist might have access only to the intensity part of the output field rather than the entire complex (intensity and phase) part.

In this thesis, I study both inference and control using the optical light field propagation in a multimode fiber . I also show that the developped framework can be applied in neuroscience to learn the transformation of a optical image to a set of electrical spikes in the vision system of the retina. I discuss challenges of dealing with real-world systems such as data collection, robustness and time variations and show real-time application of our proposed methods.

1.1 Inference in physical systems

Reconstructing the inputs of a physical system from measurements of its sensory outputs is a common practice in various disciplines such as neuroscience [4], microscopy [1], healthcare, among others. Depending on the number of measurements and partial/full observation of the system's states, various methods for recovering the original inputs of the system have been proposed. For linear and time-invariant systems, the system could be probed with many inputs and the measuring the resulting outputs. The collected input-output data then could be used to solve a set of equations to obtain the transfer function of the system in the matrix form whose successful estimation requires full observation of the system's outputs. The latter is often very expensive and requires non-trivial sensory apparatus. In cases where obtaining data is expensive or the system is nonlinear and/or is only partially observed, other methods need to be sought. For example, compressed sensing has been adopted for recovery of inputs in underdetermined systems such as in FMRI where obtaining a large data set is impractical [5]. Deep learning methods for inference is one technique that has recently been widely used in many applications such as optical imaging in scattering media [6] [7] [8] [9] [3] [10] [11] [12] [13]. In this framework, end-to-end deep learning methods have been intensively applied for information retrieval from partial measurements. In particular, the partially measured output lacks a portion of data (for example phase information) but are still able to deduce the input of the system.

1.2 Control in physical systems

In physical system characterization, a fundamental challenge is finding the proper continuous space input to a system that yields a desired functional output. For example, an open question in sensory/motor neuroscience is how to determine the input stimulation able to induce a desired behavior. Another challenging and open problem is to control the output of an optical system, such as a turbid medium used for imaging, that could be non-linear and time-varying. In a linear physical system, the problem of finding the input that produces a

desired output can be determined by monitoring its response to a series of arbitrary inputs and then computing the inverse of the system's transmission matrix (a mapping from inputs to outputs). This entails measuring the responses of the system fully. In practice, physical systems can only be partially measured and, more importantly, are often nonlinear. So, the linear transmission matrix formalism cannot be used. Even though the forward path of the system could be fully characterized, obtaining its inverse for large scale systems involving millions of variables is computationally intensive if not entirely intractable. Resorting to data-driven methods that do not require full-measurements or linear approximation of the system, such as deep learning approaches, have been shown to be successful. Deep learning techniques proposed for these tasks [14], [15] mostly take advantage of labeled data to do supervised training. For applications that require control over the response of one or an ensemble of targets, end-to-end supervised learning can fail due to the lack of labeled data within the distribution of desired target responses as well as inherent sensitivity of supervised approaches to perturbations in out-of-training-distribution data.

1.3 Thesis organization

The organization of the thesis as well as a brief overview of the chapters is summarized below.

1.3.1 Chapter 2: inference

I start off by studying the inference for the phase retrieval problem with a sub-Gaussian measurement matrix. In particular, I use multimode fibers (MMFs), which is a medium akin to a complex scattering medium for which the transmission matrix can be experimentally obtained. The phase retrieval problem seeks to find the input information of the system (the information is complex valued) from intensity-only measurements (the output information is similarly complex valued but I assume that the experimentalist has access only to the amplitude information). The cases of 1. amplitude output from amplitude input and 2. amplitude output from phase input is investigated. The inference is done in a maximum-likelihood setting using deep neural networks. Two different network architectures are used. Ultimately, extensions of our results to other inference problems in the field is discussed.

1.3.2 Chapter 3: control

In this chapter, I tackle the challenge of controlling the output of a given system when only partial measurements from the system is available to the experimentalist. I introduce a maximum-likelihood model for learning the forward transfer function of the system. A second network is then trained jointly with the forward model estimator that generates the required input of the system for producing a desired target output. I test the two-network algorithm on data obtained with the multimode fiber experimental framework. I showcase the success of the algorithm for real-time projection of arbitrary images through MMFs. Robustness of

the algorithm against perturbations is discussed. A probabilistic version of the proposed algorithm is also proposed.

1.3.3 Chapter 4: vision

The probabilistic model introduced in the previous chapter is used to control the spiking activities of Retinal Ganglion Cells (RGCs) in the vision system of subjects such as salamander/rat. I investigate if the input-output relationship (visual stimuli and spike activities) in this system could be encoded in a small number of variables (latent variables) using Information-Bottleneck formalism. With the proposed method, I show that the complexity of the input stimuli required for producing RGCs' spike activities can be substantially reduced while spike responses remain highly correlated with the original spike activities elicited by the original high resolution stimuli.

1.3.4 Chapter 5: Conclusion and future work

Finally, I conclude the thesis and provide insights for future work.

2 Inference in the scattering media

Some of the material presented in this chapter can be found in the following papers:

- B. Rahmani, D. Loterie, G. Konstantinou, D. Psaltis, and C. Moser, "Multimode optical fiber transmission with a deep learning network", Light: Science Applications, vol. 7,no. 1, pp. 1–11, 20.
- E. Kakkava, B. Rahmani, N. Borhani, U. Tegin, D. Loterie, G. Konstantinou, C. Moser,and D. Psaltis, “Imaging through multimode fibers using deep learning: the effects of intensity versus holographic recording of the speckle pattern”,Optical Fiber Technology,vol. 52, p. 101 985, 20.

2.1 Multimode fiber characterization

Multimode fibers (MMF) were initially developed to transmit digital information encoded in the time domain. There were few attempts in the late 1960’s and 70’s to transmit analog images through MMF [16] [17] using holographic recording in materials. With the availability of digital spatial modulators and cameras using digital holography, practical image transfer through MMFs has the potential to revolutionize medical endoscopy. Because of the fiber’s ability to transmit multiple spatial modes of light simultaneously, MMFs could, in principle, replace the millimeters-thick bundles of fibers currently used in endoscopes with a single fiber only a few hundred microns thick. That, in turn, could potentially open up new, less invasive forms of endoscopy to perform high-resolution imaging of tissues out of reach of current conventional endoscopes. Methods of imaging in multimode fibers (MMFs) involves measuring the phase and amplitude of the electromagnetic wave, coming out of the MMF and using these measurements to infer the relationship between the input and the output of the MMF. Most notable techniques include analog phase conjugation [17] [18] [19] [20] digital phase conjugation [21] [22] or the transmission matrix method. The latter technique, which is the current gold standard, measures both the amplitude and phase of the output

patterns corresponding to multiple input patterns to construct a matrix of complex numbers relaying the input to the output [23] [24] [25] [26]. This matrix is then used for imaging of the output or projection of desired patterns. Other techniques rely on iteratively optimizing the pixel value of the input image to perform a particular task (such as focusing or displaying an image) [27] [28] [29] [30] [31]. The dependence of the aforementioned methods on the phase measurement is also their weakness. This is rooted in two reasons. First is the necessity of having a non-trivial phase measurement apparatus. A holographic experiment requires an external reference beam brought to the output of the fiber to generate an interference pattern from which the complex optical field (amplitude and phase) can be extracted. Although some work has shown that the reference beam can also be sent through the same MMF [32], multiple quadrature phase measurements must be done to extract the phase, making the process computationally intensive.

The second reason is the sensitivity of the phase to external perturbations. Any mechanical variation or thermal variability, among others, could drift the phase of the reference wave. Upon significant change of the phase, the calibration process needs to be repeated. Therefore, careful phase tracking needs to be implemented to correct for phase drift, which further complicates the implementation. Thereby, a method that can characterize the MMF without using the phase information of the output wave while at the same time is as general as the gold standard methods is highly desired. Notably, some works have used convex optimization to infer the matrix from intensity measurement only [33] [34]. Although these works are promising steps for phase-independent characterization of the MMF, they lack generalization. For example, only a limited types of images, mostly sparse, could be imaged through the fibers.

Recently, data driven methods have been applied for characterizing scattering media and MMFs. These techniques rely on inferring the statistical characteristics of light propagation through the MMF system through examples. In what follows, I explain the first proposed data driven work with MMFs for imaging in detail and show that this approach simplifies considerably the measurement system and experiments while being able to correct external perturbations.

2.1.1 Optical fibers basics

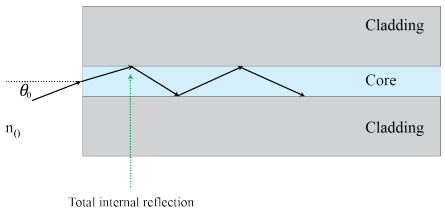

Figure 2.1 – Guiding mechanism in optical fibers.

Optical fibers, schematically depicted in 2.1 are an example of waveguides that allow light waves propagate through them. They have a concentric layered structure with an inner core

cylinder made of a higher refractive n_{core} index material (usually silica) that is sandwiched by a cladding cylinder of a lower refractive index n_{cladding}. The difference between the two refractive indices is the core mechanism that guides lights through the optical fiber. The mechanism, known as total internal reflection, allows any incoming light impinging on the input facet of the fiber to couple to the fiber provided that the incident angle θ_{inc} is lower than a predefined angle known as critical angle θ_{c}. This critical angle depends on the properties as follows:

$$n_0 \sin\theta_{\text{inc}} = \sqrt{n_{\text{core}}^2 - n_{\text{cladding}}^2} \tag{2.1}$$

where n_0 is the refractive index of the incident light medium. The right hand side of Eq. 2.1 is known as the numerical aperture (NA) of the fiber. The light that couples to the fiber is then decomposed into a number of waves with particular waveforms known as modes. Formally, a mode is a pattern of light that maintains a constant shape as it travels through the fiber. Each mode propagates with its own velocity that depends on the properties of the fiber and the solutions of the fundamental Maxwell equations satisfying the boundary conditions in the fiber. Single mode fiber allows only one mode to propagate while in MMFs several modes can propagate at the same time. Simultaneous propagation of several modes is the main advantage of the MMFs compared to the single mode fibers as the former has the capacity of transmitting more information. For example, an image that is made of thousands of pixels could directly be transmitted by a multimode fiber whereas in single mode fibers, the pixels should be sent through the fiber one at a time. Despite this superiority, MMFs are difficult to work with in practice. This is due a phenomenon known as modal scrambling. As stated earlier, an incoming wave (say a 2-dimensional image) that couples into the fiber is decomposed into the several modes. Each mode propagates with its own propagation constant (or equivalently with different phase velocity). The decomposed modes reach the output facet of the fiber accumulating their own phase differences. Accordingly, the output field of the fiber becomes distorted. With sufficient knowledge, the input wave could still be reconstructed.

2.2 Problem setting

Learning-based methods for imaging through MMFs seek to retrieve the input information (usually a 2D image) entering the system from intensity-only measurements of the output. In particular, as the phase information of the complex wave exiting the distal facet of the fiber is lost due to the squared-law of the detector (a CCD or CMOS camera), these methods seek to reconstruct the input from statistical characteristics of the system learned from data. It should be noted that such a problem is highly ill-posed as many inputs can result in the same amplitude profile at the output of the fiber that only differ in their respective phase information.

Accordingly, the information retrieval task (generally referred to in the literature as *phase retrieval*) is formally formulated as follows. Find the complex input vector of the system,

$\mathbf{X}^* = \{x_i^*\} \in \mathbb{C}^n$, that corresponds to the output, $\mathbf{Y}^* = \{y_\mu^*\} \in \mathbb{R}^m$, given the partial measurements of the system as in $y_\mu^* = \left|\sum_i F_{\mu i} x_i^*\right|^2$, where x_i^* (and respectively y_μ^*) are elements of the input (output) vector and $F_{\mu i}$ is the complex-value measurement matrix.

The optimization problem can then be re-written as:

$$\mathscr{L} = \min_{\zeta} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \Big[D(\mathbf{Y} - M_\zeta(\mathbf{X})) \Big] \tag{2.2}$$

where M_ζ is the mapping function parameterized by ζ that retrieves the information and D is a metric that measures the similarity of the predicted information to the ground-truth. For D, I either use a 2D Pearson coefficient-based similarity metric, i.e. $D = -\log[(1+\sigma)/2]$ where σ is the 2D Pearson coefficient or a l_2 norm Mean Squared Error (MSE). By construction, the observed system's outputs, y_μ^*, are always positive real values. However, the inputs are complex in general. I consider two cases where the inputs are either phase-only, $x_i^* = e^{\varphi}$ where $\varphi \in \mathbb{R}$, or amplitude-only, $x_i^* \in \mathbb{R}$.

2.3 Spatial light modulator (SLM)

To modulate the input field entering the system (here the MMF), I have used a liquid crystal phase-only SLM [35]. As the name suggests, the SLM can only modulate the phase of the incoming light. Applying an electrical voltage to each pixel of the device results in the rotation of the birefringent crystals in that pixel producing a particular refractive index change and hence a phase difference for light in that pixel. Upon reflection, the light gets modulated by a desired phase adjusted on the device. If the device is to be used for complex modulation (both amplitude and phase), the desired complex field needs to be preprocessed, i.e. the initial complex field is first mapped into another field with phase-only information (unit amplitude). This second field then produces the desired complex field only after propagation of the light. In the literature, a number of methods have been proposed for implementing this preprocessing step with various optical efficiencies [28] [30] [36]. Gerchberg-Saxton (GS) is the chosen algorithm throughout the experiments conducted in this thesis. Before starting the algorithm, the complex field is Fourier transformed and the Fourier components of the field within the numerical aperture of the MMF is retained. The GS algorithm is conducted in four steps. (1) The amplitude of the complex field is hard set to unit value (phase-only constraint) and the resulting field is then Fourier transformed. (2) The components of the new field that lie within the numerical aperture of the fiber is then replaced with the components of the original field. (3) The resulting field is then inverse Fourier transformed into the spatial domain. (4) The amplitude of the new field in the spatial domain is again hard set to unity and all the steps are repeated for a number of times (50 iterations for example). The final field is the sought-after phase-only pattern that produces the same field as that of the initial complex

pattern after propagation into the far field.

2.4 A note on the dimensions of the measurement matrix F

The measurement matrix F can be measured in multiple basis domains including pixel domain, MMF's modal domain or in frequency domain [26] [37]. In frequency domain, to obtain the output y_μ^*, one needs to take the Fourier($\mathscr{F}$) transform of x_i^*, multiply the result by F and then take the inverse Fourier transform ($\mathscr{F}^{-1}$). SLM images entering the MMF are naturally low-pass filtered due to the limited bandwidth of the MMF (spatial frequency upper bounded by the maximum k vector). As such, with a MMF of numerical aperture 0.22, I obtained a measurement matrix $F \in \mathbb{C}^{51^2 \times 51^2}$.

2.5 Experiments

The network architectures and optimization scheme is further explained in the Appendices A. I used various dataset as the inputs of the system: EMNIST dataset [38], ImageNet [39] as well as data randomly selected from a distribution such as the uniform distribution. Examples of dataset are depicted in Fig. 2.2. All images are sent through the MMF as depicted in Fig.2.3 to obtain the output speckle patterns.

In what follows, I report the performance of the information retrieval for amplitude-only/phase-only inputs for the above dataset and discuss my observations.

Figure 2.2 – Dataset examples for the reconstruction task in MMF.

Amplitude-only input: in this case, the information is entirely encoded into the amplitude of the system's input with a global phase constant chosen for all inputs. Hence the system's input is of the form $x_i^* = a_i^* e^{\varphi_0}$ where a_i^* contains the encoded information. Although the inference is hard-constraint by the network's architecture to produce predicted inputs with a global phase constant zero ($\varphi_0 = 0$) (the network output is by construction real-valued; hence zero phase), we note that any other global constant phase is also the solution of the inference. I trained my networks on the Latin alphabet dataset from EMNIST (60000 training, 1000 test data). Inference results for some sample images from in-distribution (test dataset) and out-of-distribution dataset (non EMNIST images) is depicted In Fig. 2.4 and 2.5, respectively.

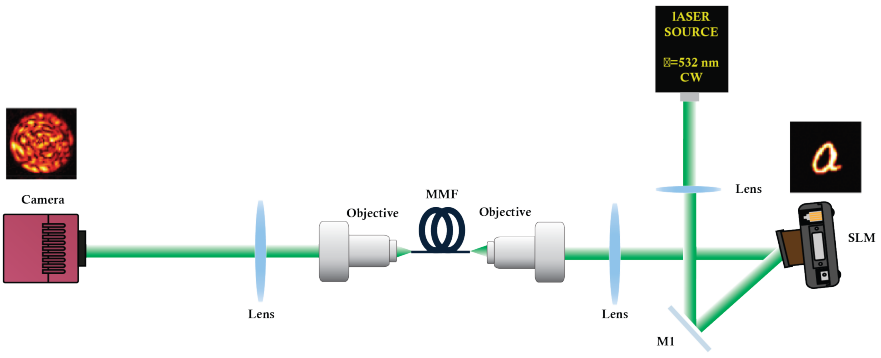

Figure 2.3 – Apparatus for obtaining the MMF output for a given input pattern.

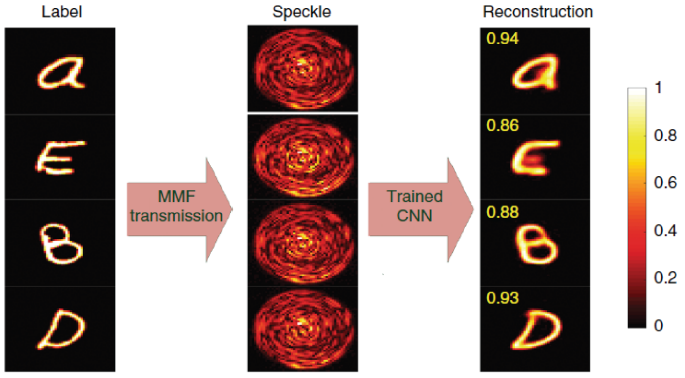

Figure 2.4 – Examples of the output amplitude speckle patterns and the reconstructed fiber input amplitude patterns produced via the CNN. The fidelity number for each reconstructed image with respect to its corresponding grayscale label is shown.

Phase-only input: contrary to the previous scenario, the information here is entirely encoded into the phase of the system's input with a global amplitude constant chosen for all inputs. Hence the system's input is of the form $x_i^* = a_0 e^{\varphi_i^*}$. Although the inference is hard-constraint by the network's architecture to produce predicted inputs with a global amplitude constant unity ($a_0 = 0$), we note that any other global constant amplitude is also the solution of the inference. As the previous case, I trained my networks on the Latin alphabet dataset from EMNIST (60000 training, 1000 test data). Inference results for some sample images from in-distribution and out-of-distribution dataset is depicted In Fig. 2.6 and 2.7, respectively.

Fully-connected vs. convolutional networks: the inference in the previous section was carried out using convolutional-type neural networks. I hypothesized the same inference could be implemented with fully-connected neural networks as the underlying physical system can be approximated with a complex-value transmission matrix as seen in the full-measurement scenario. I repeated the experiment of amplitude-only information retrieval with a fully-connected neural network and Relu activation trained on natural images from

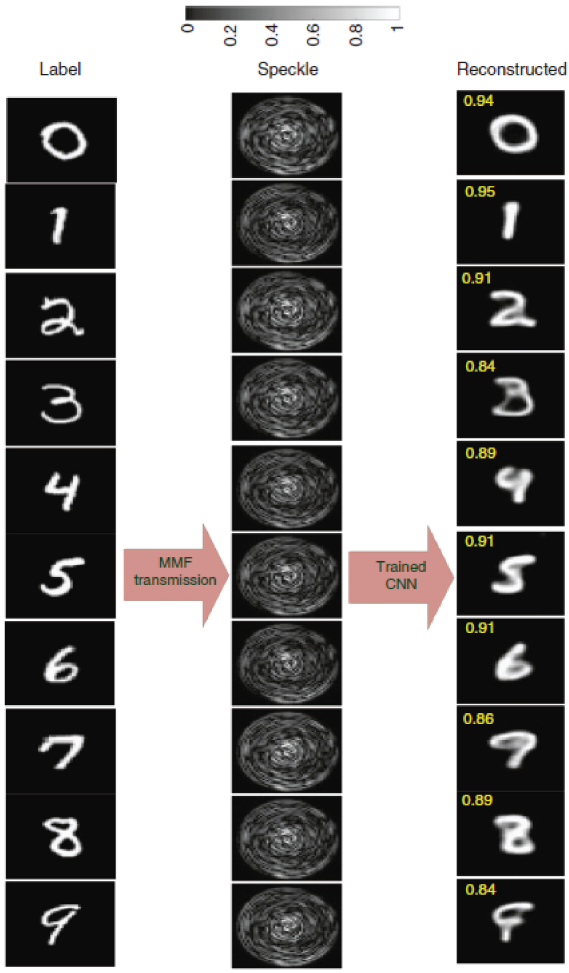

Figure 2.5 – Performance of the network in inference of out-of-distribution samples. Reconstruction of the input amplitudes from the output amplitude speckle patterns when the CNN is trained with the handwritten Latin alphabet. The speckle pattern for each image is obtained using the transmission matrix of the system.

ImageNet. As expected the, the information could still be reconstructed with a fairly good fidelity. The result of the inference on some sample images is plotted in Fig. 2.8. We note that the method here learns to relate the nonlinear amplitude-to-phase/amplitude inversion in the real domain. Thus, the network is effectively learning a sub-space, instead of the complete space, which the matrix learns, by using a much simpler measurement apparatus.

Variation of the training-data size I additionally study how the size of the training dataset could affect the inference fidelity. I use input data that are sampled separately and independently from the uniform distribution $x^* \sim U(0,1)$. I define the projection (dot product) of the reconstructed $\hat{x_i}$ onto its corresponding ground-truth x_i^*, i.e. $\hat{x_i}.x_i^*$ (vectors are normalized

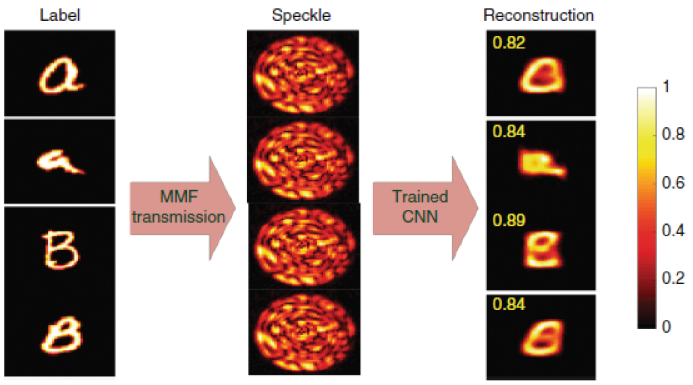

Figure 2.6 – Examples of the output amplitude speckle patterns and the reconstructed fiber input phase patterns produced via the CNN. The fidelity number for each reconstructed image with respect to its corresponding grayscale label is shown.

before dot product), as a measure of the algorithm's fidelity.

Figure 2.9 plots the MSE and reconstruction projection for various (number of training samples) 10k, 1000, 100. Note that the number of test samples is always 10 percent of the training sample number (test and train data are produced independently and from the same distribution).

2.6 Discussion

Comparing the inference fidelities of the amplitude-only and phase-only scenarios together, we notice that the former is superior to the latter. This roots from the types of nonlinearities that are present in each case. In the amplitude-only setting, the nonlinearity is due to the intensity-only detection at the distal-end of the system (fiber), where the phase information is mixed with the amplitude. In the phase-only setting, however, an additional source of nonlinearity is the exponent of the system's input. Hence, this double nonlinearity renders the inference more challenging and hence the inference fidelity is worse on average by 11 percent.

The above observation is also manifested in fidelities and convergence times of the inference carried out by different neural network architectures. The Res-net network which is equipped with more complex learning architectures than the VGG-net (such as skip connections, etc.) obtains better performance than its simpler counterpart in the phase-only inference task.

Table 2.1 compares the inference fidelities for both amplitude- and phase-only tasks obtained by both types of network architecture.

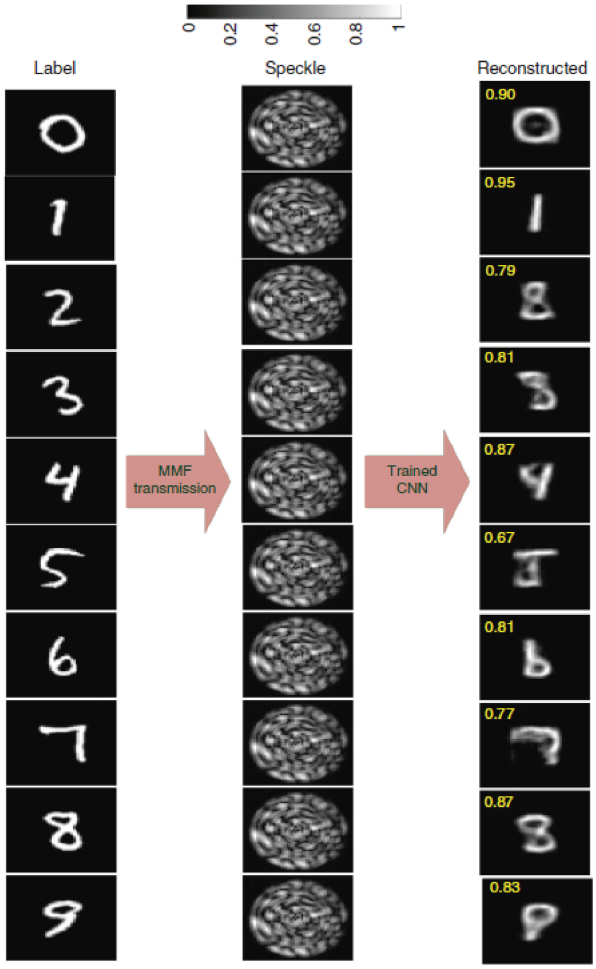

Figure 2.7 – Performance of the network in inference of out-of-distribution samples. Reconstruction of input phases from the output amplitude speckle patterns when the CNN is trained with the handwritten Latin alphabet. The speckle pattern for each image is obtained using the transmission matrix of the system.

Network architecture	amplitude-only	phase-only
VGG-net	0.93	0.79
Res-net	**0.96**	**0.88**

Table 2.1 – Inference fidelity: Phase/amplitude-only vs. network architecture

2.7 Related works and extensions

This work [15] together with [40] [41] spurred new lines of research within MMF imaging. Other authors showed the same performance of the inference in MMFs with more complex

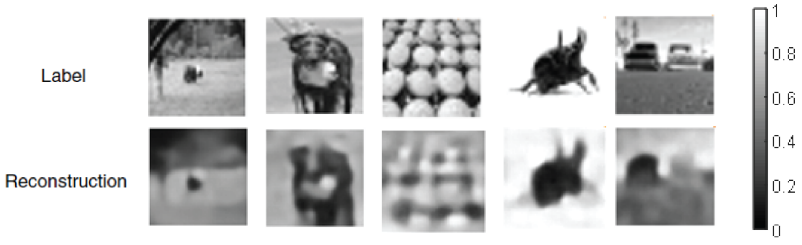

Figure 2.8 – Examples of the reconstructed amplitude images obtained from a fully-connected network trained on samples from ImageNet [39].

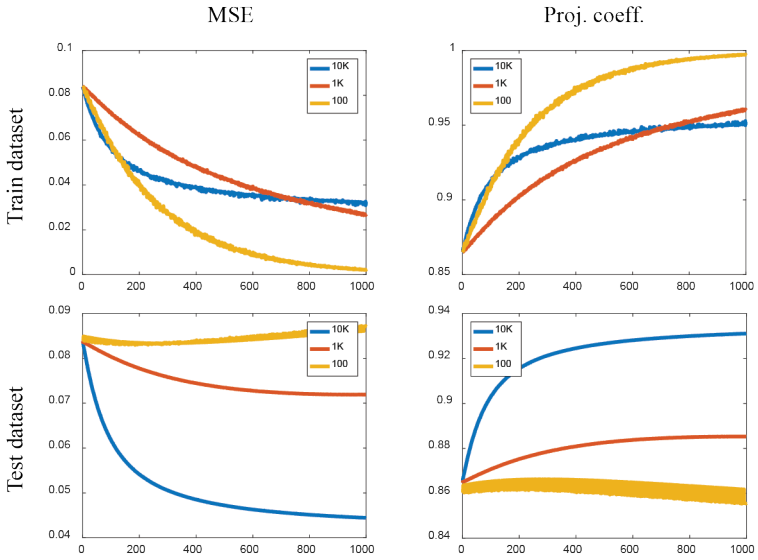

Figure 2.9 – Examples of the reconstructed amplitude images obtained from a fully-connected network trained on samples from random input.

input images [42]. Authors in [43] use simpler architecture neural networks consisting of single layer fully-connected networks for reconstruction of the images scrambled through MMF with comparable fidelity as that of the complex architecture neural networks.

One particular extension was learning the MMF system perturbations. Specifically, authors therein showed the applicability of the data-driven approaches for long lengths of MMFs up to 1 km where the system is expected to be very sensitive to external perturbations. They showed that despite the pronounced sensitivity of the fiber to external perturbation specifically in the case of the long 1 km length, it was observed that the network learns to correct for the perturbation and successfully reconstruct the information. In keeping with the previous results, authors in [44] attempted to learn perturbed system of MMFs. In particular, the fiber was moved around a number of different but fixed configurations where in each position examples of the input-output of the system were collected. A network was trained with the

entire dataset (combined dataset from multiple configurations) to reconstruct the input of the system. Even though the fiber undergoes sever change from one configuration to the other, It was shown that reconstruction of the input is still possible. This is an evident advantage of learning-based methods as compared to their non-learning counterparts; for in the latter case, as soon as the system is moved, a recalibration of the entire system is necessary. Similarly in [45], authors applied deep learning to the image retrieval problem that shows robustness to fiber deformations as large as few millimeters. By drawing from a method that combines data from different configurations of the MMF (configuration learning), images decorrelated by a factor of 10 (Pearson correlation of 0.1) because of fiber bends, were reconstructed with high fidelities. The authors attribute this success to CNNs learning invariant properties in the speckle produced for different fiber conformations. Similar methods have been applied to larger fiber bends, for example in [46] where authors show successful reconstruction of the input image for 5 cm fiber bend.

Another source of perturbation is the drift in the wavelength of the laser source that decorrelates the output intensity with time. In a series of studies conducted by Kakkava et.al. [47] [48] [49], it was shown that the DNNs can correct for the decorrelation rendered by the wavelength change of the fiber with an extended bandwidth. External perturbations that are detrimental for imaging could be harnessed for sensing in MMFs. Authors in [50] use deep learning for sensing, such as temperature for example, using the complex optical interference output of a MMF. The method is shown to work even when the information is buried in strong undesired noise. In other line of works, the spectral output of the MMF is used for sensing mechanical perturbations such as bends [51] [52] along the fiber.

3 Control of the scattering media

Some of the material presented in this chapter can be found in the following papers:

- B. Rahmani, D. Loterie, E. Kakkava, N. Borhani, U. Tegin, D. Psaltis, and C. Moser, "Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media",Nature Machine Intelligence,vol. 2, no. 7, pp. 403–410, 2.
- B. Rahmani, D. Psaltis, and C. Moser, "Variational framework for partially-measuredphysical system control", 4th workshop of Machine Learning for Physical Sciences, NeurIPS, 2021.

3.1 Motivation

Reconstructing the inputs of a physical system from measurements of its sensory outputs is a common practice in various disciplines such as microscopy [1] and optical tomography [53], among others. Most learning approaches for information retrieval such as supervised learning methods, generative networks based on Generative Adversarial Networks (GANs) [54] or Variational Autoencoders (VAEs) [55] and compressive sensing approaches [56], [57], [58] rely heavily on labeled data to train deep neural networks that can faithfully recover the original inputs of the system. The Neural Networks that reconstruct the inputs from output effectively learn the reverse path i.e. from output to input as illustrated by Fig. 3.1. A more difficult problem is finding inputs of the system that results in a desired target output. This effectively means to learn the forward path. This is a common scenario when one is dealing with the problem of controlling a system when either the system is unknown or is too complex to be modeled, a setting in which no labeled input-output data from the distribution of target outputs might be available for supervised learning, a priori. In these settings, imposing a particular prior on the solutions of the inverse problem, i.e. the mapping from partially measured output to the input of the system, can encourage solutions whose resulting outputs lie within the desired part of the system's support [59], [60], [61]. The physical laws of the

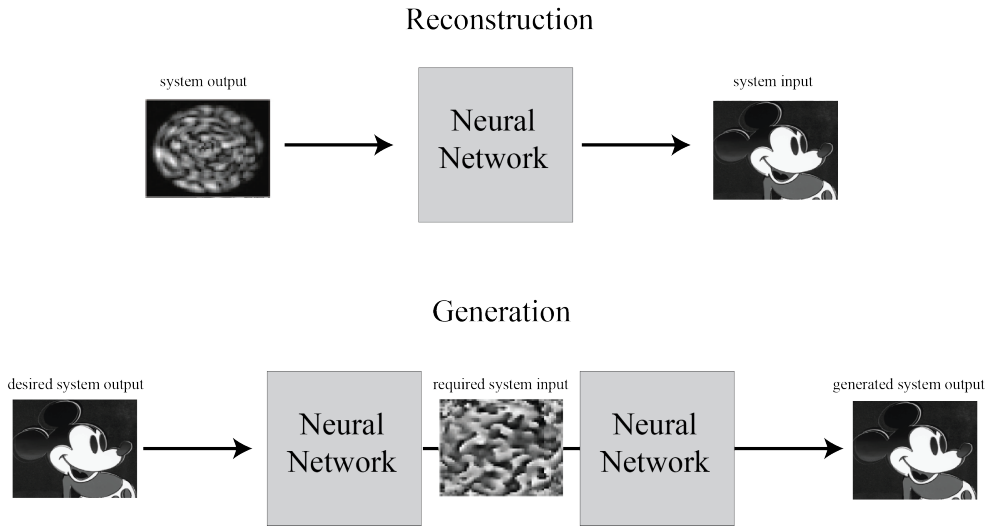

Figure 3.1 – Reconstruction and generation processes.

system could be leveraged as a prior to find solutions.

In this chapter, I assume the systems I am dealing with are theoretically constant in time. However, they are allowed to have slow variations with time due to perturbations. We will see how this affects the performance of the methods proposed.

In this chapter, I first propose a learning framework which involves the construction of a forward estimator of the system. Once the forward model is obtained, a second estimator is trained to provide the required input of the system for producing the desired output. This latter estimator could be constrained so as to promote certain solutions. At the end, I extend my work and show the proposed model is in fact a special case of the generative probabilistic models [54], specifically Variational Auto Encoders (VAEs) [55]. I compare the two frameworks and show their applicability for different scenarios.

3.2 General problem setting

In the most general form, I assume that a given input of a system, x_i, is mapped to its output via the function f as in $y_i = f(x_i)$. Therein, $x_i \in \mathbb{C}^n$ and $y_i \in \mathbb{C}^m$. The function f could be a (non)linear time-varying function. I assume that all the noise sources are incorporated in f. Additionally, f can be sampled as many times as needed. In other words, the measured output of the system, i.e. y_i, is available for any given input x_i. Furthermore, f might only be partially measured in the sense that the input-output relationship is modified by another function φ such that $\tilde{y}_i = \varphi[f(x_i)]$ where φ is either identity (fully measured system) or some other function (for example modulus $|.|^2$ function).

I seek to find the input x_i^* that would produce a desired output y_i^*. It is worth emphasizing

that the experimentalist might only have access to the partially measured system $y_i = \varphi[f(x_i)]$ while the objective is to obtain the desired output in the fully measured system $y_i = f(x_i)$. The problem, in its most general form, can be formulated to minimize an error function $\mathscr{L}$ as follows:

$$\mathscr{L} = \min_{\xi,\zeta} \mathbb{E}_{x_i,y_i,z}\Big[D[y_i, M_\zeta(x_i)]\Big] + \mathbb{E}_{y_i^*}\Big[\sigma[M_\zeta(A_\xi(y_i^*)), y_i^*]\Big] \tag{3.1}$$

where $M_\zeta : \mathbb{C}^{n\times l} \to \mathbb{C}^m$, referred to henceforth as the Model, is a differentiable representation of f parameterized by ζ and $A_\xi : \mathbb{C}^{n\times m} \to \mathbb{C}^n$, referred to henceforth as Actor, is a mapping that produces the input that feeds the Model M_ζ. The first term of the loss function represents the minimization of the distance D between sampled output y_i (experimentally) from $y_i = f(x_i)$ and the output of the Model M_ζ. This ensures that the forward model is close to the real forward propagation. The second term of Eq. 3.1 represents the minimization of the distance σ between the desired target y_i^* and the predicted output of the Model M_ζ given the output of A_ξ. The two-term loss function $\mathscr{L}$ is then jointly optimized with respect to the parameters ζ and ξ. I denote the first term in Eq. 3.1, as the Model's loss $\mathscr{L}_{M_\zeta}$ and the second term as the Actor's loss $\mathscr{L}_{A_\xi}$.

3.2.1 Forward estimator learning

The forward mapping M_ζ is maximum-likelihood estimator of the outputs given the inputs. The parameters of this function, i.e. ζ are learned in a supervised manner. The forward model's loss function in Eq.3.1 can be written as:

$$\mathscr{L}_{M_\zeta} = \min_{\zeta} \mathbb{E}_{x_i\sim\rho(x)}\Big[\log[p_\theta(y_i|x_i)]\Big] \tag{3.2}$$

where $p_\zeta(.|x_i)$ is the neural network of the forward model.

3.2.2 Phase retrieval for optical system control

The problem at hand involves controlling the forward propagation of light in a multimode fiber, which can be considered a stochastic physical system because its scrambled output is slowly time-varying. The objective is to find the appropriate complex input vector of the system (amplitude and phase of the light), $\mathbf{X}^* = \{x_i^*\} \in \mathbb{C}^n$, that produces a desired target output, $\mathbf{Y}^* = \{y_\mu^*\} \in \mathbb{R}^m$ (sampled from a desired distribution ρ, for example the distribution of MNIST dataset [38] or natural images [39]), given that the measurement of the light field is carried out with a camera which is an intensity measurement, i.e. $y_\mu^* = \left|\sum_i F_{\mu i} x_i^*\right|^2$, where x_i^* (and respectively y_μ^*) are elements of the input (output) vector and $F_{\mu i}$ is the complex-valued transmission matrix.

Although the problem is essentially a phase retrieval (PR) problem of the system's input, there are key differences with conventional PR settings: specifically, in the original PR problem the function F is entirely known *a priori*. In the current setting with MMFs, F is not directly measured and therefore is unknown. Instead, tuples of an arbitrary inputs $\mathbf{X}$ and corresponding outputs $\mathbf{Y}$ are available. Secondly, while in the conventional PR, outputs $\mathbf{Y}$ (generated via a teacher model) always belong to the support of F (the domain of all possible outputs that can be generated by function F), the target output $\mathbf{Y}^*$ may not belong to the support of F which requires finding the input that produces the closest output to the target in some metric. The optimization problem can then be re-written as:

$$\mathscr{L} = \min_{\xi,\zeta} \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left\| \mathbf{Y} - M_\zeta(\mathbf{X}) \right\|_{l_2}^2 + \mathbb{E}_{\mathbf{X}^*,\mathbf{Y}^*} \left[\| \mathbf{Y}^* - M_\zeta(A_\xi(\mathbf{Y}^*)) \|_{l_2}^2 \right] \tag{3.3}$$

where I choose the l_2 norm for the distance metrics in Eq. 3.1.

3.2.3 Training procedure

The training algorithm for the problem at hand is schematically depicted in Fig. 3.2.

Initially neither of the networks (M and A) is trained; hence I start by collecting examples for training M. These examples consist of random SLM images and their corresponding experimentally generated output speckle patterns on the camera. It should be noted that the goal here is to find a subset domain of SLM images that produce a certain class of images on the output of the fiber. In the beginning, we have no information about such a domain. Hence, in the first iteration, I choose random SLM images to send through the fiber and measure the amplitude-only images on the camera. I then, train the neural network M with this dataset. Essentially, the model M is an estimate of the forward physical light propagation in the MMF and it is estimated using random images. Once the training of M is complete, I start training the second neural network A by feeding it with the class of images that we wish to project through the fiber. The weights of the network M are kept fixed. A learns the mapping from the output amplitude images on the camera to the proper SLM images that obeys the physical propagation rules of the MMF which is modeled by M. By the end of the first iteration (i.e. training A and M), the domain of the output SLM patterns generated by network A must have moved closer in some metrics (Euclidean distance, for example) to the domain of SLM images that produce the desired images at the output of the fiber. In a second iteration, the SLM images that were produced by the network A in the previous iteration are loaded experimentally on the physical SLM and a new set of output images are captured by the camera. The network M is then retrained with this new dataset. Once completed, the training of A is carried out in the same way as in the first iteration. Depending on the quality of images that are projected through the fiber, this process could be repeated a few times. It is expected that after each iteration, the domain of the SLM patterns generated by A gets closer

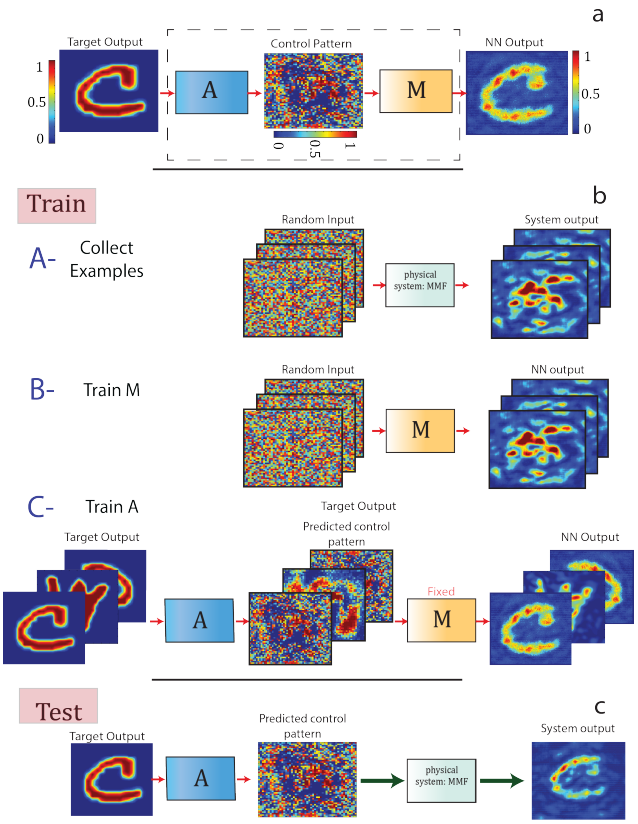

Figure 3.2 – The projector network consists of two subnetworks: the Model (M) and the Actor (A). Once trained, the subnetwork Actor accepts a target pattern desired to be projected at the output of the system (here an MMF) and accordingly generates a control pattern (here an SLM image) corresponding to the target pattern. The role of the subnetwork Model is to help the Actor come up with control patterns that are bound by the physics of light propagation through the fiber. NN, neural network. b, The training procedure is carried out in three steps. (i) A number of input control patterns are sent through the system and the corresponding outputs are captured on the camera. (ii) The subnetwork Model is trained on these images to learn the mapping from the SLM to camera, so the Model is essentially learning the optical forward path of light starting from its reflection from the SLM, propagation through the MMF and finally impinging on the camera. (iii) While the sub-network Model is being fixed (to back propagate the error), the Actor is fed with a target image and is asked to produce an SLM image corresponding to that target image. The Actor-produced SLM image is then passed to the fixed subnetwork Model now mimicking the fiber. The error between the output of the Model and the target image is backpropagated via the Model to the Actor to update its trainable weights and biases. c, The test procedure is carried out by feeding the target image to the trained subnetwork Actor and acquiring the appropriate SLM image corresponding to that target image and sending it through the system.

to the domain of SLM images that produce the desired target images at the output of the fiber. We note that the training of A, i.e. mapping from amplitude patterns to SLM images, cannot

be straightforwardly carried out because no label (ground truth SLM images) for target output images exists a-priori (training cannot be performed in a supervised manner in which ground truth labels are available beforehand). Therefore, the performance of A gets better by working synergistically with M to generate SLM images that result in output amplitude images with higher fidelities.

3.3 Experiments

I apply the proposed method of section 3.2 to obtain the required input of the MMF system for producing the target outputs. I first assume the inputs $\mathbf{X}^*$ are purely real-valued (constrain solutions to be amplitude-only). It is worthwhile to visualize a random input and its corresponding output. Fig. 3.3 depicts such examples.

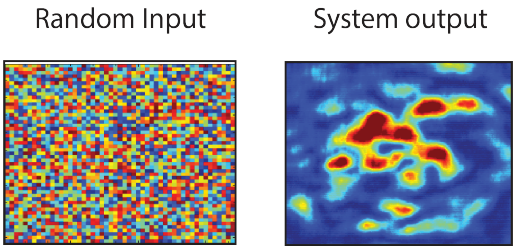

Figure 3.3 – An example of a random input to the system and its corresponding output.

I seek amplitude-only solutions to produce the target outputs of the system. Fig. 3.4 shows examples of projected images at the output (on the camera). The fidelity plot in Fig. 3.5 demonstrates the algorithm's convergence speed in finding an appropriate solution for the system.

Next I loosen the amplitude-only constraint on the sought-after solutions and seek inputs with both amplitude and phase information. This requires changing the architecture of the network. Details of the network architecture and optimization scheme is further explained in the Appendix B.

I train the networks with 20000 greyscale images of handwritten Latin alphabet characters from EMNIST as targets. Figure 3.6 shows examples of experimental outputs of the system obtained by sending the algorithm-found solutions to the system. For the sake of comparison, outputs obtained by the gold standard transmission matrix method [26], an example of methods that require full complex field measurement and control of phase information, are also shown. Without any fine-tuning, the network trained with Latin characters is used directly to project a different class of images. Examples of these projected images are shown in Fig. 3.7. These results demonstrate the generalization ability of the neural network and show that it can extend its ill-posed inverse problem ability to images never seen by the network in the training step. Comparing the projection performance of the amplitude-only and complex control patterns shows a roughly 10 percent improvement.

The performance of the neural network in inferring the required input is correlated with the complexity of the target images with which the network is trained with. The Latin alphabet images, used for training the network are of sparse nature, having a constant zero background and a greyscale feature centered in the middle of the image. Therefore, it is expected that target images with richer contexts will be more challenging. For example, one iteration was enough to project Latin images using complex modulation while the more complex images in Fig. 3.8 needed several iterations (on average 5). I explored this by using my approach to project continuous grey-scale natural-scene-like pictures. Examples of these are shown in Fig. 3.8. The projected images (for three colors corresponding to three wavelengths red, green, blue, three-channel RGB and the superposition of all three as one channel) are also depicted. The complexity in the target images makes the training difficult, but the two networks M and A are able to provide the appropriate input for producing images of natural scenes with fidelities on par with those of full-measurement schemes.

Table 3.1 summarizes the projection fidelity of the neural network approach and that of the transmission matrix approach for various types of target images.

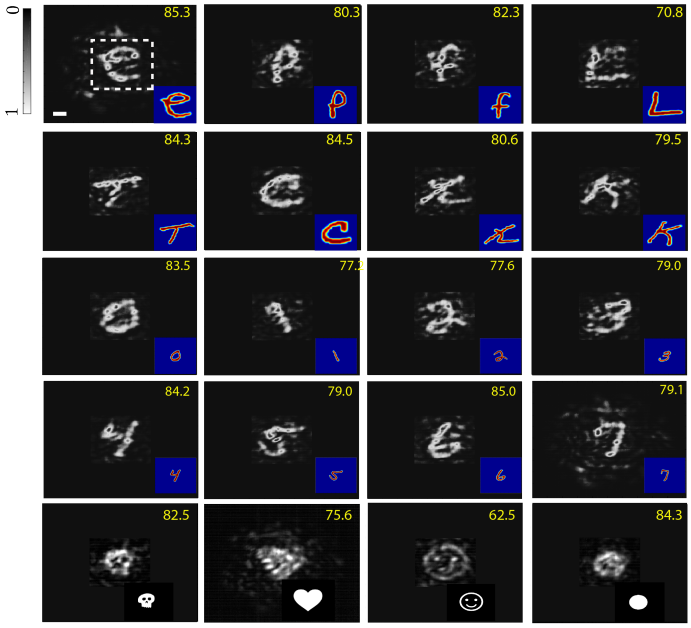

Figure 3.4 – Examples of images projected onto a camera at the output of a MMF (wavelength 780 nm) are shown. The network is forced to generate amplitude-only control patterns. These patterns are then sent to the system and the outputs on the camera are captured. The network is trained with target images of Latin characters but it is also used to predict control patterns for target images from different categories. The visible background of the projected images accounts for the lower signal to noise ratio of the images (also lower fidelities) as compared with that of the complex value control patterns. This is attributed to missing out on controlling the phase of control signals.

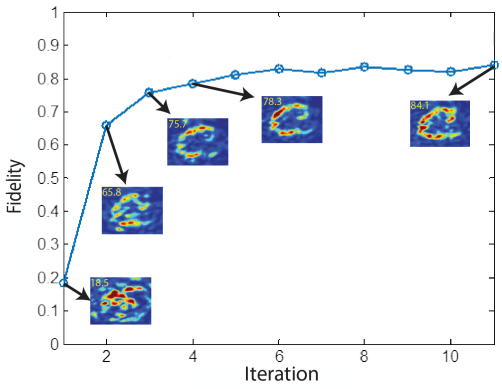

Figure 3.5 – Plot of the convergence speed for amplitude-only input controls in Fig. 3.4

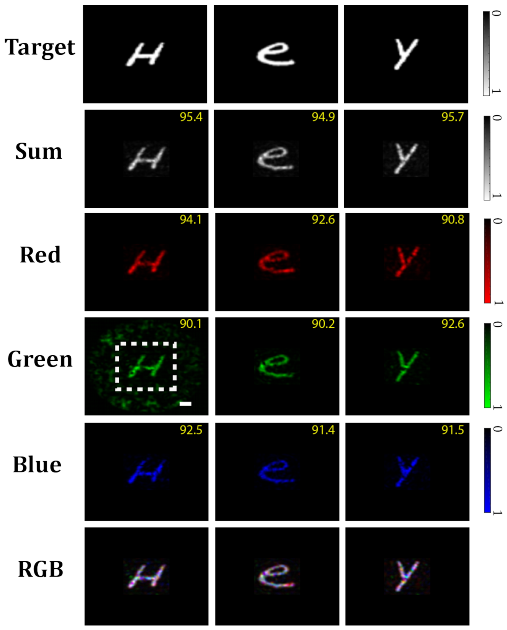

Figure 3.6 – Examples of images projected onto a camera at the output of an MMF are shown. Projection of images is carried out for three different wavelengths (633, 532 and 488 nm) corresponding to red (R), green (G) and blue (B), as well as the superposition of those colours either as a three-channel RGB image or as a one-channel incoherent image produced by summing R, G and B. The neural network is trained with the EMNIST dataset as target images. The appropriate SLM patterns generated by the network are sent to the system to obtain the desired targets on a rectangular area (200 × 200 pixels) on the camera (corresponding to an area of 19 × 19 μm^2 on the output facet of the fiber). This area is shown as a dashed box on one of the examples. Scale bar, 5 μm. The fidelities of the projected images with respect to the corresponding target images are shown. The apparatus for this experiment is depicted in Fig. B.1. We note that each color was trained and tested separately.

3.4 Extension of the forward model to probabilistic models

The forward model proposed in the previous sections was trained to maximize the probability of the outputs given the inputs. Another useful training framework is to learn a latent

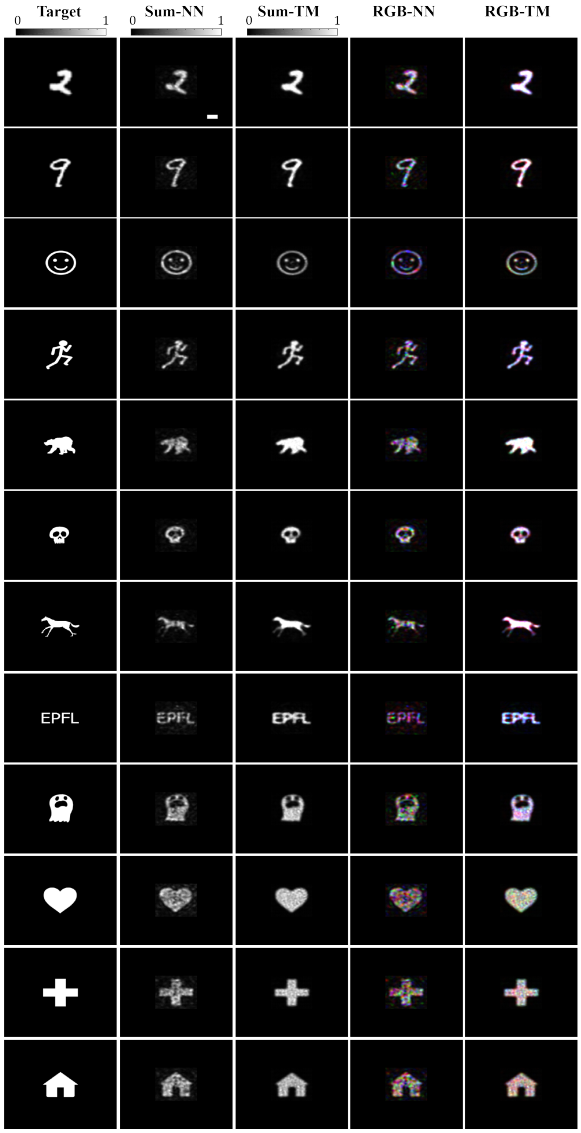

Figure 3.7 – Examples of images projected onto a camera at the output of an MMF are shown. The control patterns that produce the output images on the camera (the incoherent summation of red, green and blue wavelengths as well as the three-channel RGB images) are generated either via a neural network trained on the dataset of Latin alphabet characters (different from the category of target images) or via the transmission matrix full measurement approach (TM). The generalization of the network is demonstrated in its ability to provide control patterns for target images that come from a different class to that of the images originally used for training. Scale bar, 5 μm.

representation of the inputs and outputs data that relates the input x_i to the output y_i by: $x_i \rightarrow z_i \rightarrow y_i$ [62]. This representation may be useful as the underlying rule governing the system's propagation is captured by the latent representation in a lower dimensional space. For example, if the input-output data are noisy, the latent learning framework provides better

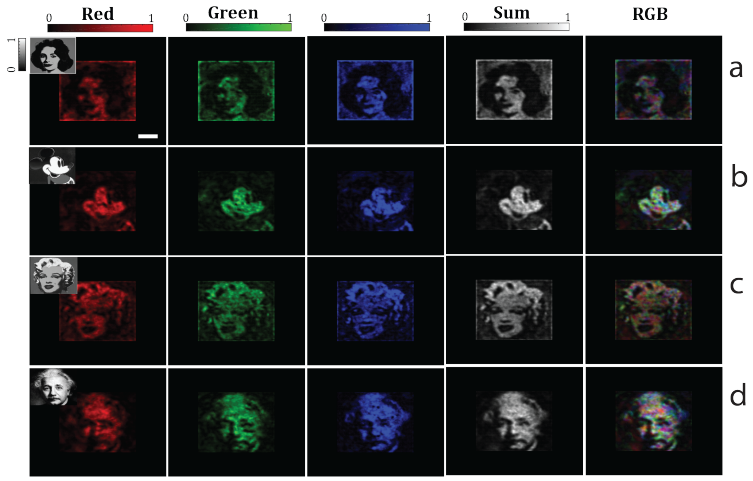

Figure 3.8 – Continuous grey-scale image projection. Examples of natural-scene continuous grey-scale target and experimentally projected images being sent through the MMF and captured on the camera for colours red, green, blue and the three-channel RGB as well as the superposition of all three colours in one channel (sum) are shown. a, Liz (Elizabeth) Taylor—1964. b, Mickey Mouse—1981. c, Marilyn Monroe—1967. d, Portrait of Albert Einstein. Scale bar, 5 mm. Credit: a–c, Andy Warhol Foundation for the Visual Arts, Inc./2020, ProLitteris, Zurich; d, Bachrach/Getty Images

Dataset	NN	TM	Avg-NN	Avg-TM	Var-NN	Var-TM
Latin-alphabet	0.95	0.981	0.924	0.969	3.7	0.8
Digits	0.952	0.982	0.925	0.971	3.5	0.7
Random sketches	0.866	0.839	0.903	0.971	3.9	1.2

Table 3.1 – Neural network and transmission matrix image projection average fidelities for various datasets. Avg-NN and Var-NN denote the mean and variance of the projection correlation using the neural network method. Avg-TM and Var-TM denote the mean and variance of the projection correlation using the holographic transmission matrix method [26].

predictions as noise cannot be fitted to the latent space and is rejected by the model [63]. In what follows, I first explain how to construct such a model and then use the developed model for the MMF dataset.

3.4.1 Probabilistic forward estimator

The forward model M_ζ introduced in section 3.2.1 can be modified based on the generative probabilistic VAE models. The reason for this choice of model is two-fold. First, forward models that are fundamentally stochastic in nature (see example 2 in Results section) could be better represented by a probabilistic model rather than a ML estimator trained in a supervised learning manner. Additionally, even if f is deterministic, noise sources incorporated into

f make it stochastic in practice. Second, the generative sampling feature of VAEs could be conveniently used to demonstrate how the correct control input x_i^* (that is required to generate y_i^*) could be obtained iteratively.

The VAE M_ζ consists of two networks, an encoder and a decoder. The encoder is trained to transform input x_i conditioned on the system's output y_i onto the latent vector z that is enforced to be close to a normal distribution of zero mean and standard deviation unity $\mathcal{N}(0,\mathrm{I})$. That is to say, each variable z_i in the latent space has zero mean and standard deviation one. The weights of the encoder network effectively learns the probabilistic conditional distribution q: $q_\Phi(z|y_i)$ parameterized by Φ. The decoder, on the other hand, takes the latent vector z drawn from the normal distribution parameterized by the encoder outputs μ_{enc} and σ_{enc}, i.e. $\mathcal{N}(\mu_{enc},\sigma_{enc})$ using reparameterization trick ($z = \mu_{enc} + \epsilon \odot \sigma_{enc}$ where ϵ is a sample from normal distribution with zero mean and unit variance) [63]- to generate the output $\hat{y}_i$; effectively learning the conditional distribution $p_\theta(y_i|z,x_i)$ parameterized by θ. The training of the VAE that models the forward Model M_ζ is carried out by optimizing the following loss function w.r.t. $\zeta : \{\theta,\phi\}$ [63].

$$\mathscr{L}_{M_\zeta} = \min_{\zeta:\{\theta,\phi\}} -\mathbb{E}_{x_i\sim\rho(x)}\Big[\mathbb{E}_{z\sim q_\phi(.|y_i)}\left[\log[p_\theta(y_i|z,x_i)]\right] - \beta\,\mathbb{E}_{y_i\sim\rho(.|x_i)}\left[D_{\mathrm{KL}}(q_\phi(z|y_i)\,||\,\mathcal{N}(0,\mathrm{I}))\right]\Big] \quad (3.4)$$

where β is the weighting factor between the two terms in the loss function, ρ is to denote a general purpose probability distribution and KL is the Kullback–Leibler divergence, a measure of how one probability distribution is different from a second one [64]. Note that, in the first term of Eq. 3.4, the logarithm of the probability density function is used because it is easier to compute the derivative of a sum than a product (log of a product is the sum of the log).

We note that setting β to zero in the forward model's loss function is equivalent to letting go of the optimization over the posteriors' latent variables which effectively becomes a maximum likelihood estimator of the forward model. When beta is not zero, the second term in the loss function can be treated as a regularization. To minimize the loss, the second term needs to be as small as possible (KL divergence is always a positive number). Enforcing the latent space to be normally distributed with zero mean and standard deviation one is equivalent to minimizing the KL function. Therefore, z solutions from the normal distribution with zero mean and unit standard variance are encouraged. On the other hand, these solutions cannot produce the ground truth outputs, resulting in a large loss from the first term. Therefore, the model opts for solutions that are close to solutions from a normal distribution with zero mean and unit variance to the extent that the output could be accurately predicted by the input. This allows the model to learn a simplified (low-dimensional) representation of the inputs that is necessary for predicting the outputs.

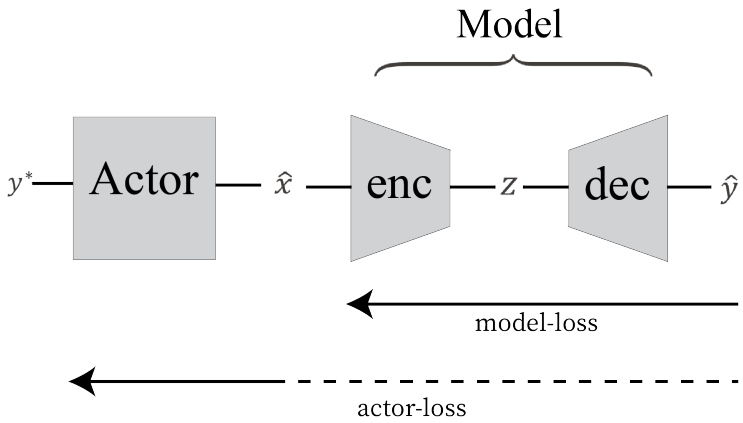

Figure 3.9 – Forward model architecture. In the figure, the target output is denoted by y^*

3.4.2 Training algorithm

A sketch of the networks and gradient flows is depicted in Fig. 3.9. Algorithm 1 in Appendix B presents the learning procedure for the system control in mathematical form. The training is similar to that of the Maximum-likelihood in Fig. 3.2 and involves in computing the variational updates of the forward model followed by training the backward model.

I again apply this algorithm on the problem of finding the appropriate input for controlling the output of the randomly scattering media of MMFs.

3.4.3 VAE forward model for phase retrieval control

I tested my algorithm with target outputs $\mathbf{Y}^*$ sampled from MNIST dataset [38]. I used 20000 EMNIST samples for training and 1000 for test. Fig. 3.10 plots the empirical l_2 norm as well as 2D Pearson correlation between the system's outputs and targets versus the iteration number I in Algorithm 1.

It can be seen that the algorithm reaches a 2D correlation of 0.83 which is close to the value (~ 0.9) obtained with the gold-standard full-measurement techniques such as described in [26] and around the same value obtained by the Maximum-likelihood model reported in Fig. 3.5. We note that this is the upper bound of the problem because in [26], the experimentalist has access to full information, i.e. phase and amplitude of the system's output (the same problem but without the modulus $|.|^2$).

3.5 Discussion

Fully-variational vs. Maximum-likelihood Comparison of the results of amplitude-only input control task carried out by the fully-variational method in Fig. 3.10 and maximum-likelihood one in Fig. 3.5 shows that both approaches perform fairly similarly. Therefore,

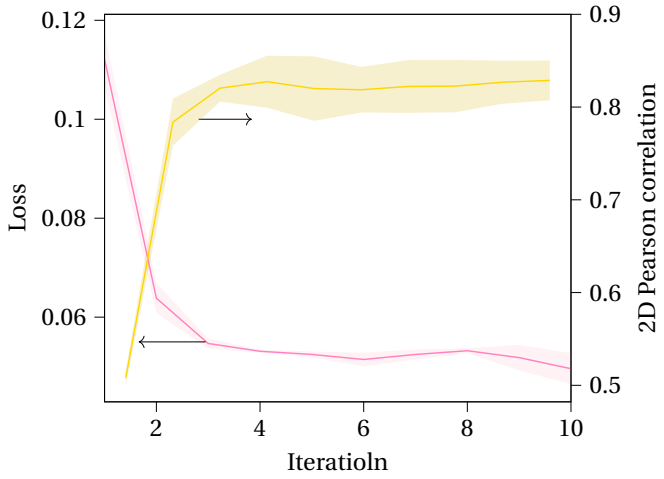

Figure 3.10 – Performance metric of the algorithm (loss: left axis and Pearson correlation: right axis) versus iteration number for phase retrieval task. The shades show the standard deviation of the results in a three-fold repetition of the experiment.

for this particular task, perhaps the maximum-likelihood method is advantageous over its counterpart as the former is less computationally intensive. The benefit of learning latent variables of a physical system becomes apparent when they are directly used to infer particular features of the system required used some tasks or the data are severely corrupted by noise. I left the analysis of the latent-variables in the MMF's forward model for future work.

Convergence speed The convergence speed of the proposed algorithms depends on the complexity of the target images, the modulation scheme (amplitude-only or complex value) and the extent by which this modulation can be implemented experimentally, and finally the rate at which the system changes over time. As it is shown in Fig. 3.5, it can be seen that the number of iterations required to achieve a certain fidelity is higher when an amplitude-only solution is used. However, the fidelity of the amplitude-only solutions are lower that that of the complex value. This is attributed to missing out on controlling the phase of control signals.

Robustness The MMF system is also prone to multiple time-dependent processes including mechanical perturbations, temperature change, instabilities associated with drifts in optical power and of the source laser wavelength, among others, which influence the learning trajectory. In a first step to investigate the robustness of the projector algorithm, I use the measured transmission matrix of the system (measured once) to virtually project the control patterns provided by the algorithm. Doing this, we are able to obtain the resulting fiber's outputs without sending them directly through the fiber. This is useful as it allows us to have a fixed replica of a experimental system that is not perturbed by external noises and is stable in time. The projector network is then trained, as before, with this new dataset. After training, the network is run to produce the SLM patterns that correspond to a user-defined output. The

patterns are then loaded onto the SLM of the experimental system. The fidelity trajectory of the projected images is shown in Fig. 3.11c (solid lines) for three colours (red, green, blue). It can be inferred from the plots that the fidelities of the virtually projected images converge to slightly higher values than those of the experimentally projected images (filled circles Fig. 3.11 c). The lower fidelity of the latter is because of the degradation due to time variation and non-perfect modulation scheme. The transmission matrix used for projecting the SLM patterns was also remeasured after each round of training to take into account the system's variation with time (dashed lines). Hence, although the system is changing over time, it is effectively being corrected. Interestingly, the close overlap between the trajectory of graphs (dashed lines) and the experimentally projected images (filled circles) shows that the neural network approach is automatically compensating drifts but without the need to continuously measure and invert the matrix, as is required in the transmission matrix approach.

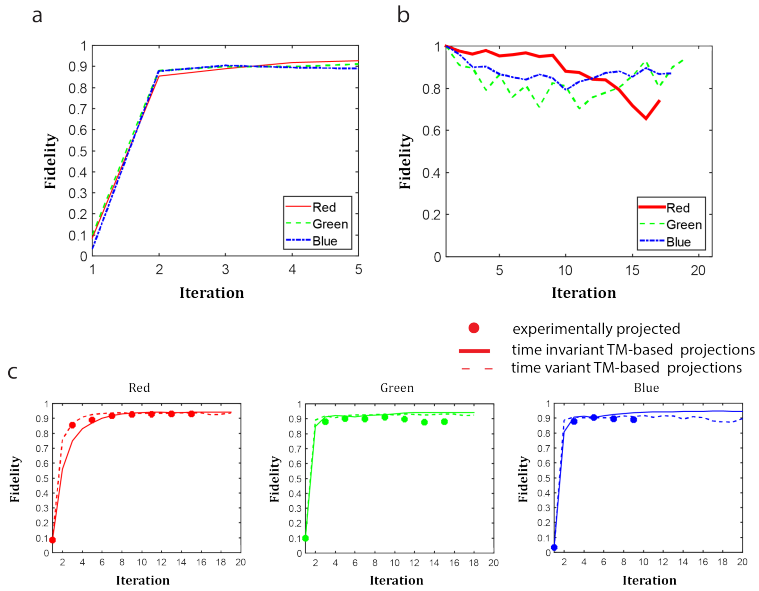

Figure 3.11 – a, the fidelity trajectory of experimentally projected images versus the training iteration number is plotted for all three colors. b, while training, the instability of the system (estimated as the correlation between instances of the system's response to a constant input signal being sent through the system over and over) is monitored over time (If the system is time-invariant, then the correlation plot holds a value of one continually). c, degradation in the fidelity of projected images due to the non-perfect modulation scheme as well as the variation of the system with time is shown by using the experimentally measured transmission matrix (TM) to forward the neural network's predicted SLM images for all three colors. The fidelities in part (a) are redrawn in part (c) for comparison. As observed, the experimentally projected images (solid circles) closely follow the track of time variant TM-based relayed projections (dashed lines) and both eventually fall below the track of time-invariant TM-based relayed projections (solid lines). In the former, what is taken out from the learning algorithm is only the effect of modulation scheme, whereas in the latter, it is the lumped effect of time variation as well as the modulation scheme. The ripples in the trajectory of the graphs in (c) (dashed lines) show that the network is continuously trying to correct for the drifts.

3.6 Related works

As opposed to the inference problem of estimating the input of the system from noisy sensory outputs in experimental disciplines such as microscopy [1], optical tomography [53] and neuroscience [65] which is a fairly well-established technique, learning methods for control applications in these fields have yet to mature. Closed-loop techniques based on deep networks have been proposed for a number of applications, such as for brain neuroscience [66] wherein authors control the activity of individual neuronal sites in V4 area by optimizing single input stimuli. Likewise, for optical turbid-medium imaging, authors have used ML-based estimators for controlling the optical fields [67]. As opposed to the previous works, I propose joint learning of the forward and backward models of the system with VAEs to implicitly impose compatibility of the sought-after solutions with the underlying physics of the problem. The latter, in essence, is akin to technique of untrained neural networks [59], [60], [61] in denoising.

4 Information bottleneck of the vision system

Some of the material presented in this chapter can be found in the following papers:

- B. Rahmani, D. Psaltis, and C. Moser, "Variational framework for partially-measuredphysical system control", 4th workshop of Machine Learning for Physical Sciences, NeurIPS, 2021.
- As of this moment, another paper from materials presented in this chapter is in preparation.

4.1 Introduction

A fundamental challenge in neuroengineering is finding the proper input to a sensory or motor system that yields a desired functional output. In an unimpaired system, this is achieved naturally by the underlying physiological circuit. For example, the vision system is an interconnected network of cells starting from light detection and transduction at the photoreceptors in the retina all the way to the visual cortex. This system, depicted in details in Fig. 4.1, is extremely complex and nonlinear. When an image is formed onto the photoreceptors, the electrical neural activity is processed by several layers of neurons within the retina and a train of electrical neural spikes is sent to the visual cortex via the optic nerve producing a perception of the visual scene (Fig. 4.2a). A disease may break this flow of information at one point while the rest of the circuit remaining intact. In this case, prostheses could be used to restore the lost functionality. For example, when photoreceptors do not function, a visual prosthesis is used in a diseased retina to artificially stimulate neural cells in the retina. The current technology in visual prosthesis consists of an array of stimulating electrodes. The number of stimulating electrodes is ~ 4 orders of magnitude lower than the number of photoreceptors (120 million vs 10 thousand artificial electrodes). Given this fact, one can ask the question: Is it possible to reduce the complexity of the input stimuli and still obtain the same functional behavior? In other words, what is the reduced spatiotemporal stimuli that elicits the same response as that of the higher-dimensional original input? (Fig.4.2d). In this case, finding the proper input to

the system is an extremely ill-posed problem. It is a similar problematic as the control of light in a multimode fiber as seen in chapter 3.

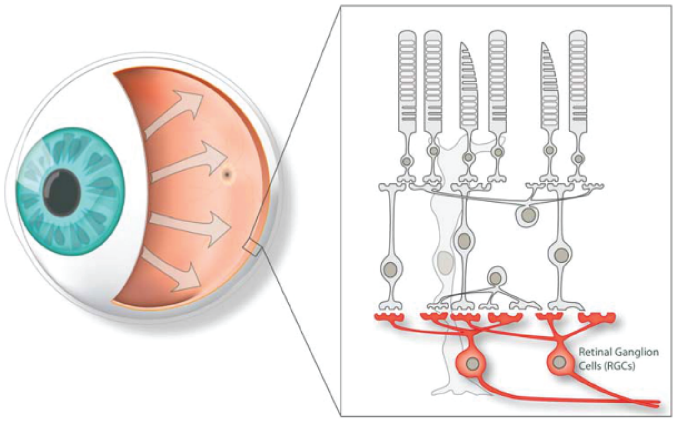

Figure 4.1 – The multi layer structure of the retina neural network. Retina Ganglion cells are located near the inner surface (the Ganglion cell layer) of the retina of the eye. Image is adopted from [68].

In this chapter, I propose a new method for characterizing the response of the retina to optical stimulation. In particular, motivated by the hypothesise about the redundancy of the information in the visual stimuli that is relevant for explaining the retina responses, I propose a model that is able to extract principal features of the complex stimuli-responses distribution that are sufficient for predicting the retina responses given the input stimuli (Fig. 4.2c). This model, inspired by the Deep Variational Information Bottleneck (DVIB) model [69], is a variational approximation to the original Information Bottleneck framework of Tishby et. al. [70]. The objective of the IB is to obtain a squeezed representation of the input source that preserves maximum information about the output response. As evident, the IB objective is conveniently aligned with that of our retina response modeling. The DVIB model assumes the IB's squeezed latent codes to be independent and identically distributed. This is obviously a poor assumption for the retina dataset that consists of time-series input images and output responses. Modeling the data sample correlations in time may provide a better model. Gaussian process priors are a good candidate for modeling the data's time correlations which happen to be conveniently integrable with the IB model. Accordingly, I propose a variational model based on the IB that uses a Gaussian Process prior to model the retina dataset. I subsequently use this model and present an adaptive stimulation algorithm which employs the prior data and the abstract representation of the IB latent space to generate redundancy-reduced stimuli that elicits output responses that are highly correlated with the original responses elicited by the complex input stimuli.

My contributions in this work are:

- **A probabilistic model**: I propose a new model for the Retinal Ganglion Cells (RGCs) spike

train in response to the complex natural visual stimuli using the IB framework. The model integrates a Gaussian process prior into the IB latent code that can learn temporal dynamics of the data in a lower dimension space. This model outperforms the state-of-the-art models of the retina responses.

- **closed loop stimuli optimization**: Using the proposed model allows only principal features of the stimuli to pass the bottleneck, the original input stimuli is pruned so that it contains the minimum information required for producing the RGC responses originally elicited by the complex input stimuli. The resulting reduced-dimension optimized stimuli is used for the next round of measurements in an iterative process.

IB is a well-established framework that has shown to work for various applications [69]. In this work, I propose a modified version of IB by incorporating data correlation as prior using Gaussian processes (GP) IB-GP.

In the following, I start by introducing the general setting of the problem and our model in section 4.2.1. The closed loop stimuli optimization procedure is outlined in 4.2.2. Experiments are presented in section 4.3. Related works are presented in section 4.4.

4.2 Method

Figure 4.3 depicts the main structure of the retina forward model. The optical stimuli **X**, denoted on the left, are 2-dimensional images entering the network in batches. Each batch consists of T images, i.e. $X_{M,1:T}$. As seen, the images in a batch are similar and hence have a high correlation with one another. The reason for this is two-fold. First, while observing a scene, images that are captured by the retina are naturally very similar. Second, during fixation, eyes undergo dynamic movements which shift the center of the gaze. To simulate this, images used during experiment have been jittered. Images of a batch are then encoded by the *encoder* network into a lower dimensional space. For illustration purposes, in the figure, this space is made of three variables z^1, z^2, z^3. These latent variables are then decoded by the *decoder* into the retinal responses $Y_{N,1:T}$. The construct of the networks (encoder to low-dimensional space) allows us to enforce particular assumptions on the underlying biology/physics of the problem. Accordingly, we can take advantage of the low dimensional space to enforce the correlation constraint on the data. Doing this in the lower dimensional space requires much less computation than in the higher dimensional space of the original stimuli (three latent variables vs. thousands of image pixels). The correlation is constrained using GPs. In a GP model, data is assumed to be drawn from a multi-variate normal distribution with a particular mean and covariance. The covariance of this distribution reflects the properties of the data. the element $k_{i,j}$ of the covariance matrix contains the relation between the i-th and j-th data. Hence, on-diagonal elements of the covariance matrix must have the highest correlation (unity). The farther two elements are in time, the lower their correlation is. An example of such covariance matrix is provided in the Fig. 4.3.

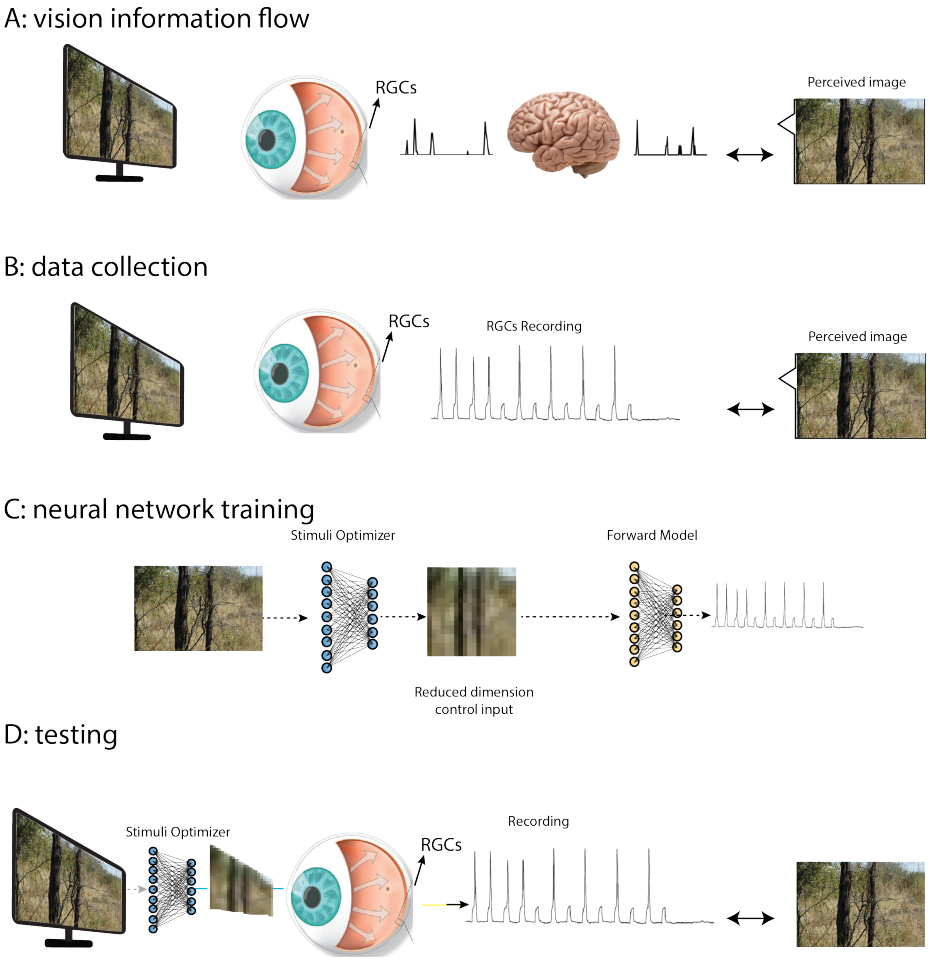

Figure 4.2 – The flow of information propagated through the visual system starts from the image impinging onto the retina. The retinal output is a spike-train at the Retinal Ganglion Cells (RGCs), which is transmitted to cortical neurons to be further processed. The model of the retinal processing will be evaluated ex-vivo with retinal explants. The end-to-end learning from the input image to the RGCs is the ex-vivo learning phase. (B, C, D) depicts the process of acquiring the required input control pattern eliciting the same neuronal activity as that of a naturally stimulated pattern. The input-output data is first collected (B) and the forward model of the system is constructed. The stimuli optimizer then explores among reduced-dimension control patterns and chooses one pattern that when fed to the Model network produces the largest correlation with the desired target spikes (C). Once trained, the optimizer network is fed with an arbitrary high-resolution image to produce a control pattern that is compatible with the resolution of the prosthetic device (D).

In what follows, I explain the mathematical derivation of the forward model's encoder and decoder and discuss the prior on the latent variables.

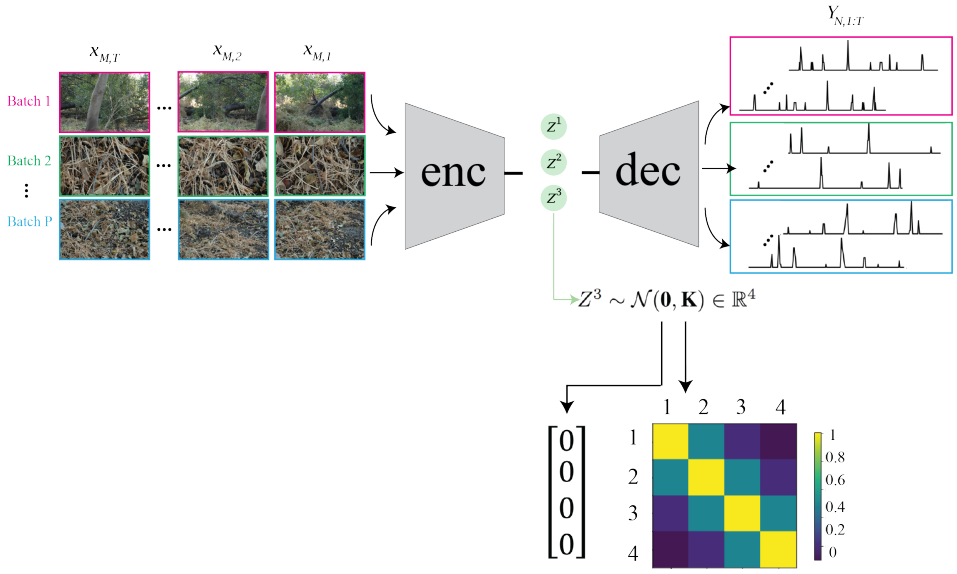

Figure 4.3 – Retina forward model structure

4.2.1 Problem setting

Input stimuli dataset $\mathbf{X} \in \mathbb{R}^{M \times T}$ consists of T consecutive spatial inputs $\mathbf{X} = [x_{M,1}, x_{M,2}, \cdots, x_{M,T}]$ that are projected in the retina at a constant rate to elicit electrical responses of the form $\mathbf{Y} \in \mathbb{N}^{N \times T}$, where $\mathbf{Y} = [Y_{N,1}, x_{N,2}, \cdots, x_{N,T}]$ and N denotes the number of RGCs.

With the above assumptions, I formulate the problem as finding the latent variables $\mathbf{Z} \in \mathbb{R}^{L \times \mathcal{T}}$ that have maximal mutual information $I(Z, Y)$ with targets $\mathbf{Y}$ in the Markov setting $\mathbf{Y}$ $\mathbf{X}$ $\mathbf{Z}$. We note that $\mathbf{Z}$ is comprised of $\mathcal{T}$ consecutive data points, i.e. $\mathbf{Z} = [z_{L,1}, z_{L,2}, \cdots, z_{L,\mathcal{T}}]$. $\mathbf{Z}$ is simultaneously constrained to have minimal mutual information with $\mathbf{X}$. Therefore, the constraint optimization problem can be written as

$$\mathscr{L}_{IB} = \max_{\zeta} [\, I_{IB}] \quad \text{where} \quad I_{IB} = I(\mathbf{Z}, \mathbf{Y}) - \beta I(\mathbf{Z}, \mathbf{X}) \,. \tag{4.1}$$

ζ denotes the parameters of the model and β is used as a trade off to adjust the amount of reduced and preserved dimensionality of $\mathbf{Z}$. To obtain these latent variables, I approximate their posterior distribution with a parametric stochastic encoder $p_\phi(\mathbf{Z}|\mathbf{X})$.

Due to the *spatio-temporal* nature of the data, assumption of data independence along the time dimension is not valid. The same is true in the latent space. To account for this fact, I assume a posterior of the form:

$$p_\phi(\mathbf{Z}|\mathbf{X}) = \prod_{l=1}^{L} \mathcal{N}(\mathbf{z}^l | \boldsymbol{\mu}_\phi^l, \boldsymbol{\Sigma}_\phi^l) \tag{4.2}$$

where $\mathbf{z}^l$ denotes the l-th latent vector of size $\mathcal{T}$. In Eq. 4.2, $\boldsymbol{\mu}_\phi^l$ and $\boldsymbol{\Sigma}_\phi^l$ are the outputs of the encoder neural network. These two outputs estimate the mean and variance of the p_ϕ distribution. I choose the structure of the covariance of the multivariate normal distribution in Eq. 4.2 so as to reflect the time correlations of the data. Hence, similar to previous work [71] [72], I construct the $\boldsymbol{\Sigma}_\phi^l$ in the model by the product of bidiagonal matrices. Intuitively, the non-zero adjacent elements of the covariance matrix (bidiagonal) should help to capture the temporal correlations of the consecutive data samples.

$$[\boldsymbol{\Sigma}_\phi^l]^{-1} = \mathbf{V}_l^{\mathrm{T}}\mathbf{V}_l + \mathbf{I} \text{ where} \tag{4.3}$$

$$[\mathbf{V}_l]_{\tau\tau\prime} = \begin{cases} v_{\tau\tau\prime}^l & \tau\prime \in \{\tau, \tau+1\} \\ 0 & \text{otherwise} \end{cases}$$

I note that the low-rank GP kernel assumption makes the computation required for drawing samples from q linear in time [71].

Expanding the IB mutual information function in Eq. 4.1 (details in the Appendix), we have:

$$I_{IB} = \int d\mathbf{Y}\, d\mathbf{Z}\; p(\mathbf{Y},\mathbf{Z})\; \log \frac{p(\mathbf{Y}|\mathbf{Z})}{p(\mathbf{Y})} - \beta \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z})\; \log \frac{p_\phi(\mathbf{Z}|\mathbf{X})}{p(\mathbf{Z})} \tag{4.4}$$

Although all terms in the RHS of Eq. 4.4 are fully defined, computing marginal distributions $p(\mathbf{Z})$ and $p(\mathbf{Y}|\mathbf{Z})$ may be intractable. I use a variational approximation for $p(\mathbf{Y}|\mathbf{Z})$ that is parameterized with θ. On the other hand, the prior on the latent variables, i.e. $p(\mathbf{Z})$, is modeled using GPs. In particular, I assume a GP prior on $\mathbf{Z}$ defined as a multivariate normal:

$$\rho(\mathbf{Z}) = \prod_{l=1}^{L} \mathcal{N}(\mathbf{z}^l|\mathbf{0},\mathbf{K}) \tag{4.5}$$

where ρ is the variational approximation for the prior and $\mathbf{K}$ is the covariance function that models temporal covariances in the latent space. In more details, the covariance between τ-th and $\tau\prime$-th samples is computed as $\mathbf{K}_{\tau\tau\prime} = \mathcal{K}(\tau,\tau\prime)$ where $\mathcal{K}$ is the kernel function.

Substituting the variational approximations for the intractable marginals into Eq. 4.4 and using the fact that the Kullback Leibler (KL) divergence is always positive, I obtain a lower

bound for the IB objective:

$$I_{IB} \approx \frac{1}{T}\sum_{t=1}^{T}\left[\mathbb{E}_{p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))}[\log q_\theta(y_t|\mathbf{Z})] - \beta D_{KL}[p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))||\rho(\mathbf{Z})]\right]. \tag{4.6}$$

4.2.2 Closed-loop stimulation

In this section, I take advantage of the model developed in the previous sections to devise an algorithm that uses the prior recorded data to optimize the stimulation by iteration, in a closed loop (Fig. 4.4). Specifically, I assume the latent variables of the learned model, henceforth referred to as the *forward model*, have captured the principal rules governing the underlying biological system (Fig. 4.4)a. Now, I intend to find a transformed version of the original input stimuli that yield a set of latent variables that produce the most correlated responses with the original target responses in the subsequent round of measurements. Of course, the original stimuli themselves are one set of solutions that produce the highest correlation (unity) with the original responses. However, this is not a useful transformation. Instead, I am interested to obtain the best transformation subject to a constraint on their complexity. Hence, I define a parametric function $f_\xi : \mathbb{R}^{M\times T} \rightarrow \mathbb{R}^{M\times T}$ that maps the complex original stimuli $\mathbf{X}$ into the transformed stimuli $\mathbf{X}^*$. Passing $\mathbf{X}^*$ to the forward model, the parameter of the mapping functions, i.e. ξ, are optimized so as to the forward model's output responses $\mathbf{Y}^*$ are the most correlated with the original responses $\mathbf{Y}$. Denoting the original stimuli and responses as $\mathbf{X}^{\text{orig}} := \mathbf{X}$ and $\mathbf{Y}^{\text{orig}} := \mathbf{Y}$, the objective function reads as:

$$\min_{\xi} \mathscr{J}_{\text{adap}} = D(\mathbf{Y}^{\text{orig}}, \mathbf{Y}^*)$$

where:

$$\mathbf{X}^* = f_\xi(\mathbf{X}^{\text{orig}}),\ \mathbf{Z}^* \sim p_\phi(.|g(x^*_{1:\tau(t)})),\ \mathbf{Y}^* \sim q_\theta(.|\mathbf{Z}^*). \tag{4.7}$$

Therein, p_ϕ and q_θ are the encoder and decoder of the forward model trained on the prior data, D is a distance metric between the original target outputs and the predicted responses of the forward model. Finally, the optimized stimuli for subsequent measurements can be obtained by following the procedure: (1) map the original stimuli to the optimized stimuli using mapping f_ξ, (2) encode the new input $\mathbf{X}^*$ to the posterior $\mathbf{Z}^*$, (3) decode the latent variables, (4) construct the metric D and backpropagate the error to fit f_ξ, (5) once converged, use the mapping function to obtain the optimized stimuli $\mathbf{X}^*$ (6) send new inputs $\mathbf{X}^*$ to the true biological system and observe the new true system outputs, (7) to repeat the procedure, use the new dataset $\hat{\mathbf{X}}$, $\hat{\mathbf{Y}}$ to fine tune the forward model and repeat the procedure. This procedure is depicted in Fig. 4.4.

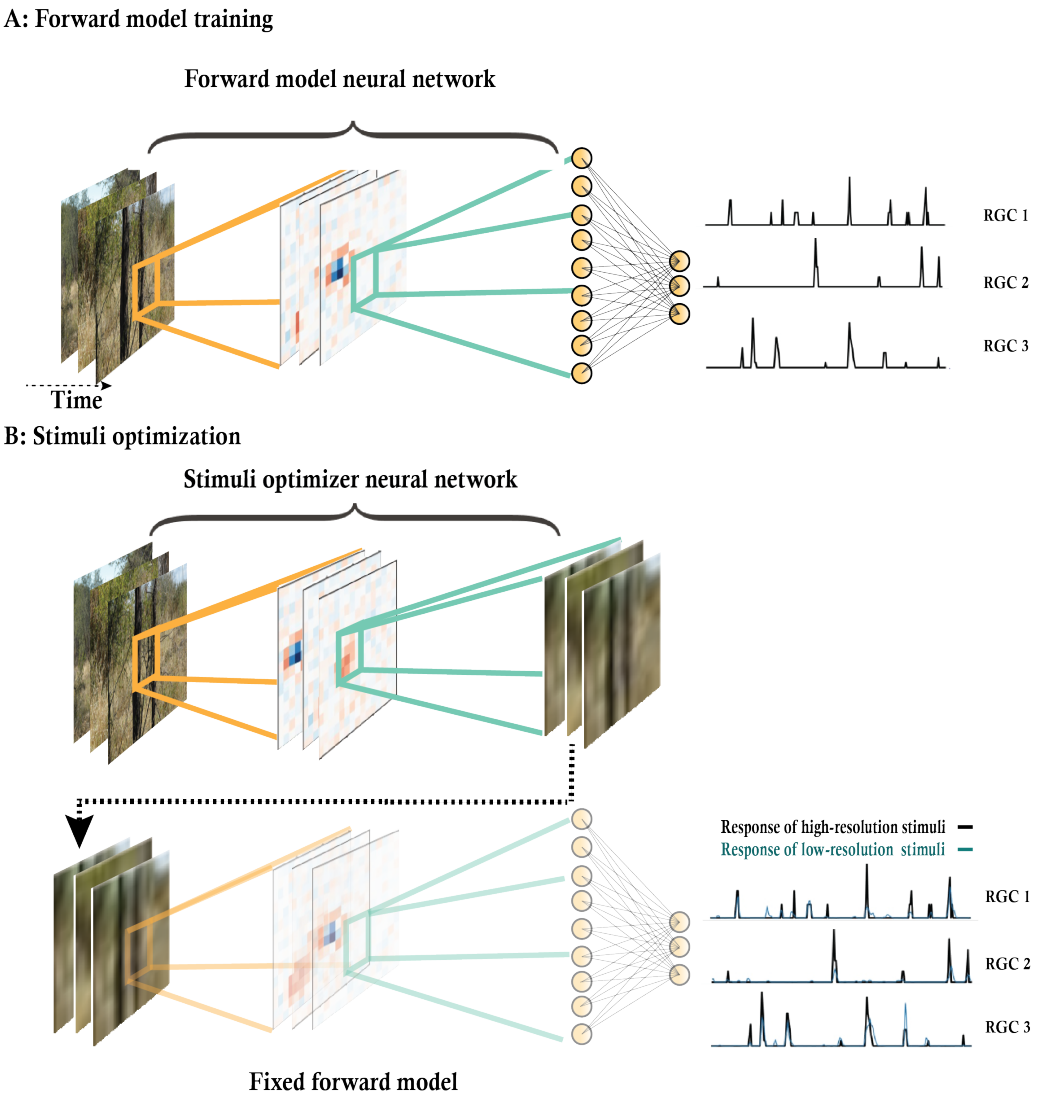

Figure 4.4 – Detailed schematic of the forward and stimuli-optimization models of the system. (A) Learning the forward mapping of the system. (B) Learning the stimuli optimization so that the reduced-dimension stimuli elicit the same responses as those of the high-dimensional ones (blue versus black RGC responses in B). Convolutional architecture of neural networks well resembles their biological counterparts (as depicted activation functions resemble On/off bipolar cells [14]).

4.3 Experiments

I focus on applying our model to the time-series dataset containing the 2D train of input stimuli (images) entering the vision circuit of an example subject and the corresponding elicited count responses of a group of neurons. Specifically, I use a real-world experimental dataset *Natural-small* which contains natural scene images as input stimuli and spike trains

Table 4.1 – Performance of various methods in the literature used for modeling of the retina network. Pearson correlation (higher better), KL (lower better), Negative Log Likelihood (NLL) (lower better)

	Natural-small			Natural-small-smooth
Model	NLL	KL	Pearson	Pearson
Feedforward CNN [14]	0.095	-	0.322	0.416
IB-Disjoint	**0.087**	3.279	**0.396**	**0.494**
IB-GP	0.089	**1.672**	0.375	0.475

from nine RGCs in Salamander [14] [I].

I start with the end-to-end modeling of the forward path of information transmission through the vision circuit. I compare our proposed model against the maximum likelihood-trained CNN baseline and show that it can consistently outperform the baseline method. I attribute the superior performance of our model to the reduced redundancy of latent variables constraint to learn only the principal features necessary for generating target outputs from source inputs.

4.3.1 Methods considered

Below, I summarize additional methods including some variations of the original proposed method that I used for the task.

- **Maximum likelihood feedforward CNN** is used as a comparison baseline. I use the same CNN network that was shown in the previous work to obtain state-of-the-art results on larger version of *Natural-small* dataset [14]. The predicted outputs of the network are the averaged maximum likelihood estimation of the targets given the inputs. To account for the variability in retinal spiking, artificial noise is injected into the model during optimization.
- **IB-Disjoint** optimizes the objective from Eq. 4.6 using the same training procedure as that of the original method. The difference lies in that the IB-Disjoint assumes latent samples are temporally independent. Accordingly an isometric Gaussian distribution is employed as the prior in Eq. 4.6. This model is akin to the standard VAE (no latent GP) in which samples in the data and latent space are assumed to be independent.

4.3.2 Forward modeling

I used the models outlined above to fit the data pairs $\mathbf{X}$ and $\mathbf{Y}$. *Natural-small* dataset contains 359000 input-output samples in the training set, of which 20% was randomly chosen for

[I] I have made my own prototype experimental apparatus for obtaining retina dataset of Rat that is still to be tested. However, all results shown in this chapter are from Salamander dataset [14].

validation. The test set contains 5000 averaged repeated trials of novel stimuli. We note that the outputs (spiking responses) were binned at 10ms. I also build a preprocessed version of the dataset in which the outputs are smoothed using a Gaussian filter with a standard deviation of 10ms.

To account for the discrete nature of the output responses (count time series), I employed a Poisson regression model for the decoder, i.e. $q_\theta(y_t|\mathbf{Z}) = \text{Po}(y_t|f_\phi(\mathbf{Z}))$ where f_ϕ is the parametric network. The architecture of networks and the training setup is explained in greater details in the Appendix C.

I report the models' normalized negative log likelihoods on the ground truth data and the normalized KL divergences between the latent posterior and the prior. I interpret the latter metric as a measure of reduced redundancy or abstraction in the data representation. Moreover, I use the Pearson correlation [73] between the ground truth data and the networks' prediction as another metric for evaluating the performance of the models. Table 1 reports on the results of our analysis.

4.3.3 Adaptive stimulation

As we saw in the previous sections, the IB framework by construction is able to learn an abstract representations of the joint distribution of the input stimuli and the target responses. These representations were in turn used to generate transformed stimuli that elicited responses fairly correlated with the responses of the original stimuli. In this section, I use the adaptive stimulation procedure to optimize the stimuli based on the feedback from previous measurements. To show the effectiveness of the proposed method, I used simulated data to allow for multiple repetitions of closed loop procedure comprising of stimuli optimization and evaluation. Details for generation of simulated data is discussed in the Appendix. I used the IB-GP model as the forward model and optimized the stimuli in a three phase closed-loop experiment.

Fig. 4.5 Plots the correlation of the responses elicited by the optimized-stimuli with those elicited by the original stimuli at each phase of the closed-loop experiment. It can be seen that the algorithm almost reaches the maximum possible performance of the system (1D correlation ~ 0.3) within three iterations. The latent vector of the forward Model (shown in Fig. 4.6) is sampled at each iteration and projected to a 2-dimensional (2D) embedding using t-SNE [74]. The true latent vector distribution required for obtaining the desired outputs is also shown (orange circles). Examples of the final (iteration 3) reduced-dimensional stimuli obtained by the proposed algorithm and the original stimuli as well as their elicited responses are shown in Figs. 4.7 and 4.8, respectively.

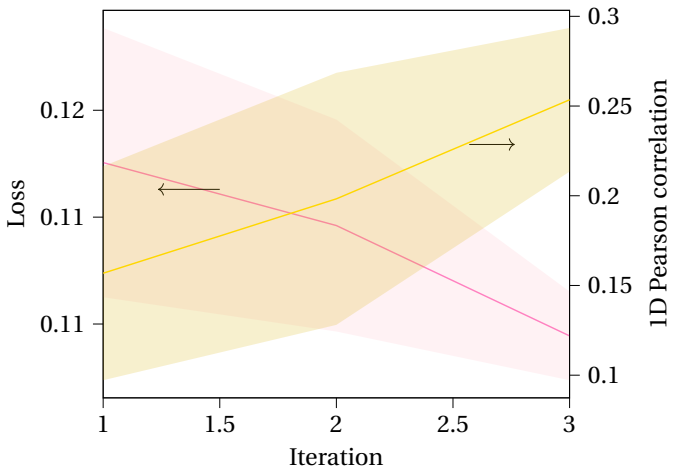

Figure 4.5 – Performance metric of the algorithm (loss: left axis and Pearson correlation: right axis) versus iteration number for closed-loop stimuli optimization task. The shades show the standard deviation of the results in a three-fold repetition of the experiment.

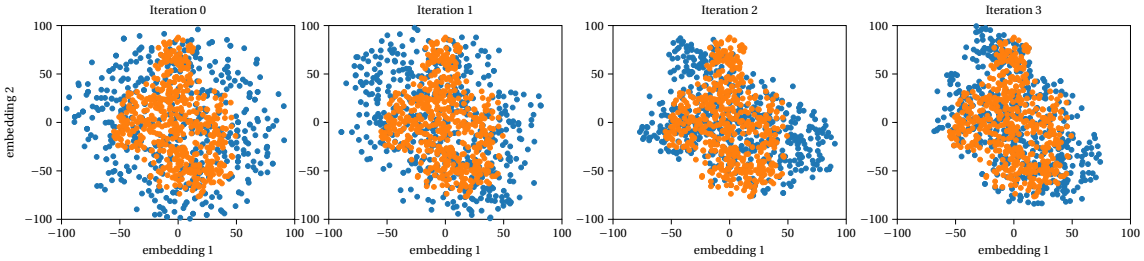

Figure 4.6 – The Latent vector evolution as 2D embedding (blue). Orange dots denote the latent vector of the true system.

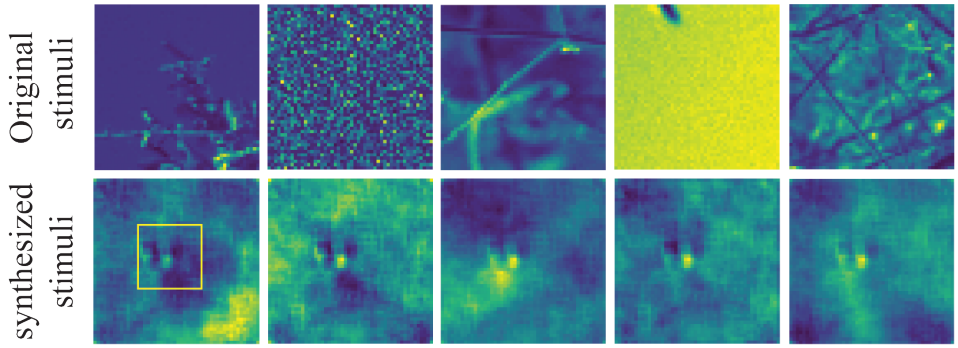

Figure 4.7 – Examples of the optimized stimuli (bottom row) and their original high-dimensional counterparts (top row). We note the significant reduction of complexity in the obtained solutions. The boxed area denotes the locations of neurons' receptive fields.

4.4 Related work

Ensemble response modeling of the retina dates back to the early work by Chichilnisky where RGCs responses to white noise was captured by averaging the stimuli inducing spikes [75]

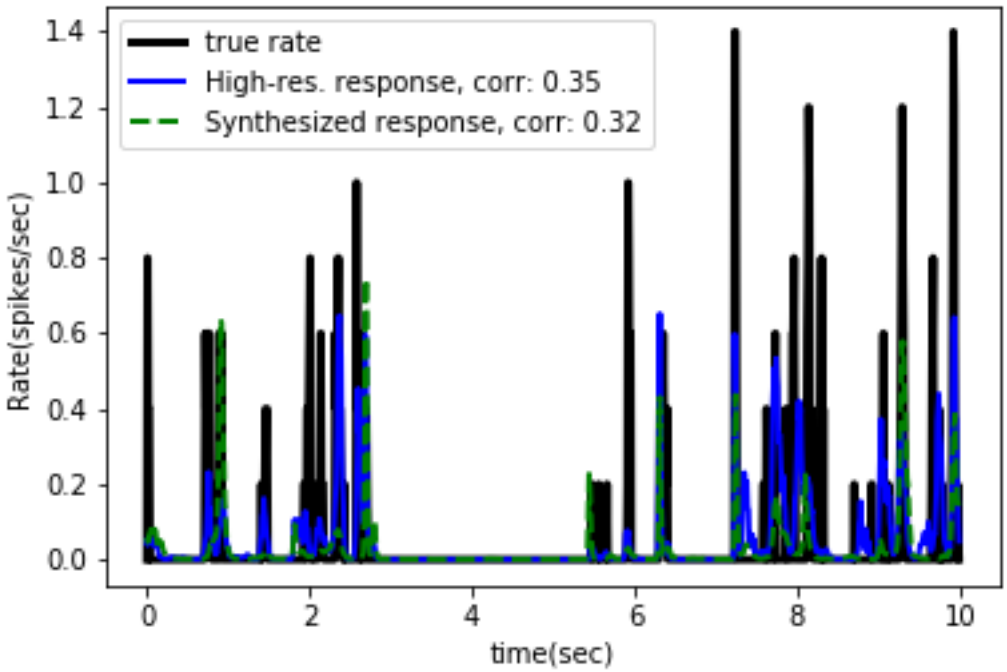

Figure 4.8 – Spike responses of an example neuron elicited by the original and optimized stimuli in Fig. 4.7

and later on by Pillow et. al. using spatiotemporal filters followed by a single nonlinear unit [76] [77]. Although successful in explaining some of the nonlinear responses of the retina, the proposed models could not explain other nonlinear behavior induced by natural visual stimuli. The first model capable of such computation was first proposed by McIntosh et. al. using convolutional neural networks (CNNs) [14] [78]. Other efforts for retina response modeling use other variants of CNNs or training formalism for example by employing recurrent neural networks [79] to fit the input-output data. On the contrary to our model which maximizes the probability of the output responses through learning of the posterior of the inputs, all models above maximize the likelihood of a neuron emitting a spike given the input stimuli in a supervised fashion (maximum-likelihood setting). To account for the stochastic nature of the retina responses, authors in [14] added raw noise to the activations of the CNN's intermediate layers. This is readily accommodated by the probabilistic latent codes of our IB model.

In the closed loop input optimization literature, our work is related to [67][80][81][66]. In particular, authors in [66] control the activity of individual neuronal sites in V4 area by optimizing single input stimuli whereas in our work, thanks to the learned latent code of the entire population of the targeted neurons, all stimuli are optimized together.

In the work of Shah et. al. [80], the RGCs' responses to electrical stimulation was optimized by first developing an model for the electrical stimuli and the spiking probabilities and then

using the model for adaptive stimulation. The model is obtained by parameterizing the spike amplitude and electrical stimulation threshold by very few parameters and then maximizing an evident lower bound on the spiking probabilities. Intended for a different application, our work does not assume any relevant relation between the input stimuli and a property of the spiking probabilities but instead, the relevant parts of the input for predicting the output is automatically discovered.

5 Conclusion and future work

This thesis presents a neural-network based learning framework for solving highly ill-posed problems to predict a system's forward and backward response functions. Such an approach has applications in inference and target-oriented system control in fields such as optics and neuroscience.

5.1 Summary of the results

The inference for the phase retrieval problem in a MMF was studied. State-of-the-art deep neural networks were used to retrieve the input information of a MMF using intensity-only measurements of the output. In particular, I used two variants of convolutional neural networks (CNNs). Residual neural networks and VGG-nets. The former network obtained average reconstruction correlation of nearly 90 percent. I saw that phase-encoded information is more difficult to reconstruct than amplitude-encoded information.

Conversely, controlling the output of a MMF was then investigated. The proposed method for this task was shown to find the appropriate continuous space input of a system that resulted in a desired output, despite the input-output relation being nonlinear, the system being slowly time-varying and with incomplete measurements of the systems variables and lack of labeled data required for supervise learning. The proposed approach consists of modeling the forward and backward propagation of the MMF using two separate neural networks that are trained jointly. The proposed method was then used to project arbitrary images through the MMF.

The proposed control learning framework was then applied to the retina network to obtain the simplest visual spatiotemporal control patterns that elicit RGCs spiking responses that are elicited by high-dimensional stimuli.

5.2 Future work

In the course of this work, several aspects of data-driven physical system learning were identified that could be the topic of future work.

Studying the latent space of the variational autoencoder (VAE) model of the MMF in chapter 3 is an interesting future work. We hypothesized that latent variables learn the governing rules of the system in a lower dimensional space. This should make the system more robust to noise. It might be that VAEs show better generalization as they interpolate in a lower dimensional space. This could be useful for imaging with memory effect without the requirement to scan all angles.

The proposed method for controlling the retina spike trains could be extended to primary cortex level. Additionally, other control methods such as Reinforcement Learning (RL) could be combined with the redundancy reduction technique of chapter 4.

Feed forward networks that are often used in the construct of forward models work better in the slowly varying settings (so that the forward model could be updated and used before the system changes again). Accordingly, I intend to study the use of memory-based RL solutions based on recurrent neural networks (RNN) and meta-RL solutions that require few data for rapid training.

Another exciting line of work is to use inverse RL (IRL) methods for better transferability and generalization. With IRL, it is possible to decouple the goal and dynamics of a control problem so that the agent could still reach its goal even when the dynamics of the environment have changed. Comparing a normal vision system with a perturbed one (e.g. when a portion of the photoreceptors are impaired), the objective is to find a policy that could still reach the goal of producing the same sensation at the cortex level as that of the normal system. In this setting, methods based on Max-Entropy and adversarial IRL are beneficial.

A Appendix for chapter 2

A.1 Experimental setup for data acquisition

The optical setup for the transmission of light through the fiber is depicted in Fig. A.1. The system here is a step-index (length = 0.75 m) MMF with a 50 μm diameter silica core and a numerical aperture of 0.22 (1055 number of fiber modes). The inputs correspond to 2D phase patterns displayed on a phase only SLM, which are then demagnified on the MMF entrance facet by the 4F system composed of lens L1 and OBJ1. The MMF output facet is imaged onto a camera. The light source is a continuous wave source at 532 nm with a power of 100 mW, which is attenuated with a variable attenuator to deliver only 1 mW for the acquisition of the images. The light source is coupled into a single mode fiber. The light beam coming out of the SMF1 (object beam) is filtered by the polarizer LP1, collimated by the lens L4 and directed on the SLM, which can spatially modulate the impinging light. The pattern created by the SLM is imaged through the relay system (lens L1 and objective lens OBJ1) at the MMF input. The quarter wave plate (QWP1) before the fiber input changes the polarization from linear to circular (this polarization is better preserved in step-index fibers). Then, light travels through the fiber and at the output, an identical relay system (OBJ2 and L2) magnifies the image of the output and projects it on the camera plane (the QWP2 converts the circular polarization back to linear) is imaged onto a camera.

A.2 Neural network architecture

Two types of convolutional neural networks, VGG-net and Res-net, have been used for the inference problem introduced in chapter 2. The architecture of the two networks is schematically shown in Figs. A.2 and A.3.

VGG-net The network consists of 12 blocks, in which the first and the last are the input and output units that are responsible to encode and decode data to the network, respectively. The input block maps the grayscale input images (one channel) to 64 channels (stack of

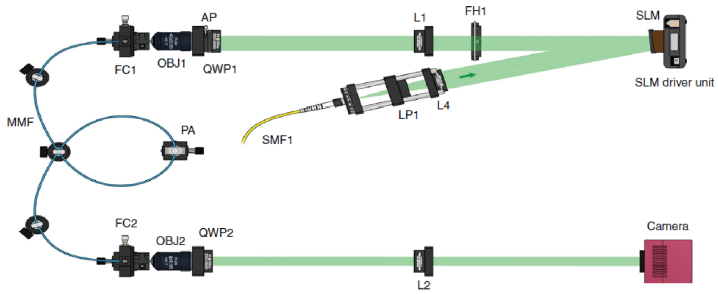

Figure A.1 – Schematic of the experimental setup for the transmission of light through the fiber. The pattern created by the SLM is imaged through the relay system (lens L1 and objective lens OBJ1) at the MMF input. An identical relay system (OBJ2 and L2) magnifies the image transmitted through the fiber and projects it on the camera plane. Image produced by Dr. Damien Loterie.

processed images) via a trainable convolutional unit. The middle 10 blocks constitute the 20 hidden layers of the neural network, wherein each block is formed by two convolutional entities that are individually followed by an element of the rectified linear unit (Relu) with a mapping functionality $\text{Relu}(x) = max(0,x)$. The convolutional layers within each block take the convolution of the stack of input feature map X_k^q using weights W_k^q and biases B_k^q complying with the formula $X_k^q = \text{Conv}_{W_k^q}(X_k^q) + B_k^q$, where the subscript k indicates the layer number and the superscript $q = \{1,2\}$ corresponds to the first and second convolution operations in each and every block. I have used only 3×3 convolution kernels. At the output of each block, an additional Max-pooling unit is considered. Max-pooling units help to avoid overfitting and therefore are essential parts of the network. These units decrease the widths and heights of the images passing through them by a factor of two. To keep the dimension of the images constant throughout the network, additional reshaping units are placed just before the Max-pooling units. These units reorder the stack of 256 $M \times M$ images into 64 $2M \times 2M$ images where M is the size of the image in both dimensions at the time of that particular operation. The Max-pooling units then downsample the $2M \times 2M$ images back to $M \times M$ images. The final block is made of a convolutional layer that simply decodes back the images from 64 channels to the original single-channel grayscale images. The architecture of the network in this work follows the standard block-structure (Conv-Relu)→(Conv-Relu)→ MaxPool, which is adopted from VGG-Nets (VGG19) [82] and customized by adding the reshaping units before the Max-pooling units.

Increasing the number of layers or the number of channels (very deep and wide architectures) adds to the complexity of the network. It is a well-known fact from the generalization theory in machine learning that more complex networks require more training data to overcome overfitting. On the other hand, the ability of the network to generalize degrades when shallow/thin networks are used. Therefore, the number of hidden layers, herein 20 for the VGG-net architecture, as well as the number of output (input) channels in the input (output) block, i.e., 64 channels, is empirically chosen based on a trade-off among the network's complexity, the

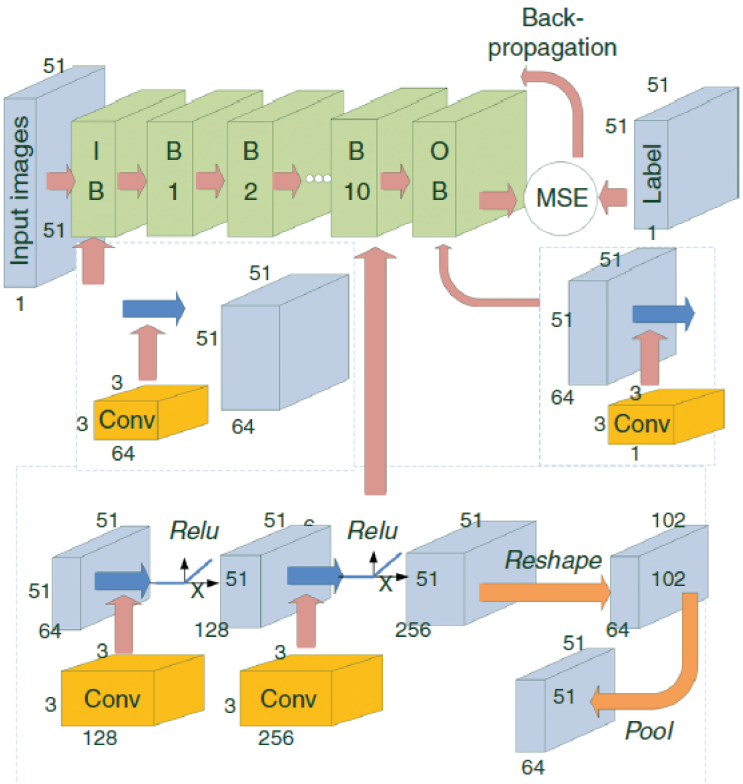

Figure A.2 – Detailed schematic of the CNN used for training and testing. IB (Input Block), OB (Output Block), Bi (Block i, where i = 1,2,3,...,10), Pool (Max-pooling), Reshape (reshaping unit). The input block maps the input images via 64 convolutional filters. Each middle block (B1-B10) contains two convolution layers followed by a reshape and max-pooling layer, which together downsample the widths and heights of the images by a factor of two. A rectified linear unit (Relu) transform is placed after each convolution unit in the hidden layers. The images are then mapped to the output channel via the convolution filters in the output block. The MSE between the labels and the processed images is then calculated and back propagated to the network to update the learnable variables.

number of training data, and the level of accuracy desired to obtain the optimal results. Additionally, more complex networks require processing units that are able to do computationally intensive calculations more rapidly. Therefore, the complexity of the network is also balanced here with the available hardware power as well as the image output time.

Res-net The architecture of the Res-net CNN is schematically shown in A.3. It consists of 9 blocks; of which, the first and the last are made of a single convolution layer followed by a non-linear rectified linear unit (Relu) which, respectively, takes one channel (64 channel) grayscale input images (output images) and maps them to 64 channels (1 channel) of stack of images. A batch-normalization [83] unit is considered after all convolutional layers throughout the network and also one at the very beginning of our architecture. I only have used 3×3-

kernel convolutional layers. The 6 blocks in the middle of the network, known hereafter as residual blocks 1 to 6, are made of two convolution layers that are both followed by Relus. The architecture of the network is based on Residual neural networks [84]; hence, a skip connection is added to the output of the second convolution layer at the end of each and every residual block and before the Relu. In the case where dimensions of feature maps after convolutional layers increase (residual block 3), extra zero entries are added to match dimensions. Max-pooling units with sizes 2×2 and strides of 2 are considered after residual blocks 1 to 4. In a similar fashion, up-sampling units are added after residual blocks 5, 6 and block 7. Note that block 7 does not have any skip connections and therefore, is not a residual bock. I have used the proposed Res-net for the task of output speckle amplitude to input amplitude and output speckle amplitude to input phase conversions similarly to what I did using VGG-net CNN. The learning rate is also similar to the learning rate of VGG-net (10^{-4}) to ensure a fair comparison in terms of the convergence rate and training time for the two architectures. Comparison between the training time and fidelity number between the two structures reveals the superiority of Residual architectures for these tasks.

Training Once the images are obtained in the final layer of the CNN (feed-forward step), they are compared with their corresponding labels in a MSE sense. The MSE in this comparison that is used for updating the learnable parameters is minimized with stochastic gradient descent. Adaptive moment estimation optimization (ADAM) algorithm [85]. To obtain accurate results within a reasonable time, I empirically choose a learning rate parameter of 10^{-4} in the optimization algorithm and a mini-batch size of 64.

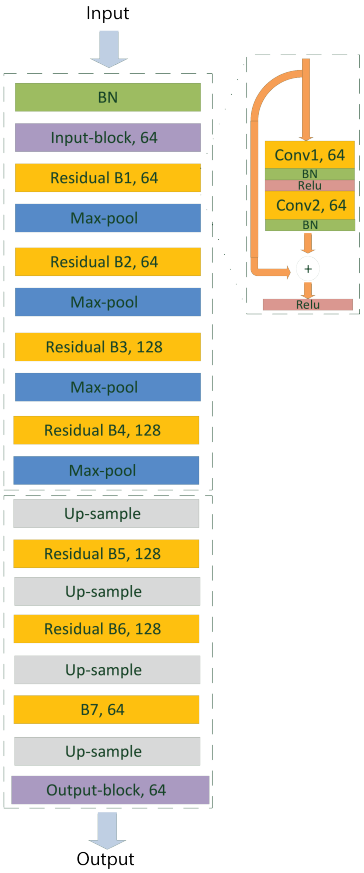

Figure A.3 – Detail schematic of the Res-net architecture.

B Appendix for chapter 3

B.1 Experimental setup for data acquisition

The experimental set-up for image transmission through the fibre is presented in Fig. B.1. Three continuous input beams at wavelengths of 488, 532 and 633 nm are delivered, one at a time, to the system via a single-mode fibre. The beam entering the system (attenuated to an average power of 4 mW) is collimated by lens L2 (f = 100 mm) and then directed to the SLM. The beam spatially modulated by the phase-only SLM (HOLOEYE PLUTO) is imaged on the input facet of an MMF using a 4–f system composed of L3 (f = 250 mm) and objective (OBJ) 1 (×60, NA = 0.85). After transmission through the graded-index fibre (length L = 75 cm, core diameter D = 50 μm and NA 0.22; corresponding to 1,050 fibre modes for one polarization), the output field is imaged onto the camera using an identical 4–f configuration. The experimental set-up for measuring the transmission matrix of the system requires the extra reference path (shown faded in the schematic). The beam in the reference path is superimposed to the main path's beam on the camera via a series of mirrors and a beam splitter. A number of input patterns (basis vectors) modulated with either phase, amplitude or both are then sent through the fibre, and their corresponding complex output fields are measured. The transmission matrix is then constructed using these input–outputs.

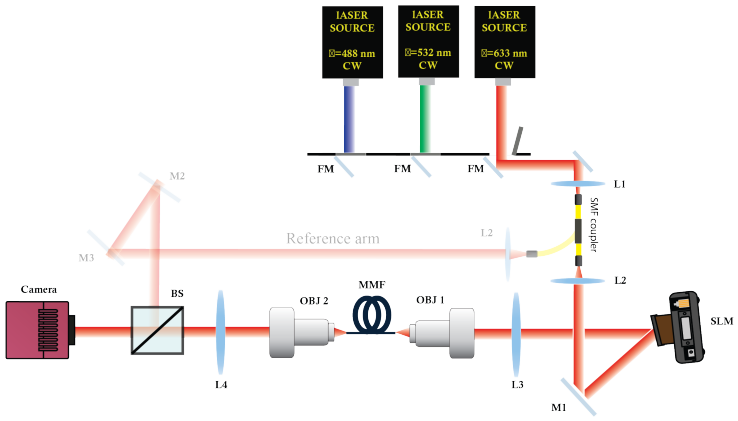

Figure B.1 – Detailed diagram of the optical setup. Control patterns are generated via the SLM, guided through the fiber and captured by the camera. L1: Aspheric lens, L2: f = 100 mm lens; L3: f = 250 mm lens; L4: f = 250 mm lens; OBJ1, OBJ2: 60x microscope objective; SLM: spatial light modulator; M1: mirror; FM: flip mirror; SMF: single mode fiber; MMF: multimode fiber, BS: beam splitter.

Algorithm 1

Input: Data tuples (x_i, y_i) sampled randomly from partially measured system $y_i = \varphi[f(x_i)]$, target outputs y_i^*, K_1 and K_2 (number of training steps for M_ζ and A_ξ, respectively), I (number of going back and forth between training the entire networks and experimenting the obtained solution in the true system)

Output: The control input x_i^* required for generating y_i^*

Initialization Variational parameters $\zeta : \{\theta, \Phi\}$ and ξ
for iter $\in \{1, 2, 3, ...I\}$ **do**

  for i $\in \{1, 2, 3, ...K_1\}$ **do**
    $\zeta \leftarrow \zeta - \alpha \nabla_\zeta \mathscr{L}_{M_\zeta}(x_i, y_i, z)$
  end for
  for i$\in \{1, 2, 3, ...K_2\}$ **do**
    $\xi \leftarrow \xi - \alpha \nabla_\xi \mathscr{L}_{A_\xi}(y_i^*)$
  end for
  Sample new (x_i, y_i) from $x_i \leftarrow \hat{x}_i = A_\xi(y_i^*)$, and $y_i \leftarrow \hat{y}_i = f(A_\xi(y_i^*))$
  Calculate empirical performance metric $\frac{1}{N}\sum_{i=1}^{N} \sigma(\hat{y}_i, y_i^*)$
  if System's desired performance is achieved **then**
    End training
  end if

end for

B.2 Variational Autoencoder training

In the training of the variational Autoencoder, first the forward model M_ζ is learned through optimizing $\mathscr{L}_{M_\zeta}$ in the general loss function in Eq. 3.1 for some number of steps K_1 (note the gradient flow in the sketch). Once done, the backward model A_ξ is learned via optimizing the loss function $\mathscr{L}_{A_\xi}$ for K_2 number of steps (gradients flow through the fixed Model to reach the Actor).

After $K_1 + K_2$ steps, I assess the performance by sampling some targets (x_i^*, y_i^*) and using the learned Actor $A_\xi(x_i^*, y_i^*)$ to obtain the system estimated inputs $\hat{x}_i$. Finally, I compute how the predicted control patterns are performing in the experimental system and reiterate the entire process this time with the solutions obtained by the backward model and their corresponding true system outputs as the inputs for training the forward model if the performance is not satisfactory.

B.3 Neural network architecture

The hyperparameters of the forward and backward networks, optimizers as well as training epochs is summarized in Table B.1 for both cases full-variational algorithm (case $\beta \neq 0$) and Maximum-likelihood version (case $\beta = 0$). Architecture of the networks in each case is also presented in Tables B.2 and B.3.

As opposed to the fully-variational case where the model network has two units encoder and decoder, for $\beta = 0$ case, I opted to use a single fully-connected unit for the M network (no latent variable is learned). Also, to allow complex-valued modulation of the inputs and outputs of the system, so as to be able to closely mimic the complex-valued physical fields entering the system and hence taking advantage of the higher degree of freedom in shaping the input fields, the Actor and Model themselves are made of two smaller sub-networks (A_{real}, A_{imag}) and (M_{real}, M_{imag}). The real and imaginary part of the A predict the real and imaginary part of the input control patterns, respectively. Each part is then sent to the corresponding real and imaginary part of the M that independently relates the real and imaginary part of the input field to the output. The training of pair M_{imag}, A_{imag} is always carried out separately, immediately after training pair M_{real}, A_{real} in an identical manner. Accordingly, I only refer to them collectively as the Actor and Model. Each of the sub-networks has the architecture of a fully connected neural network, with the Sigmoid activation.

	Fully-variational	Maximum-likelihood
Optimizer	Adam	Adam
Learning rate	10^{-4}	10^{-4}
VAE's β	500/450	-
Latent space dim.	100	-
Train/val/test batch size	20/- /1	32/-/1
Train/val/test batch num.	10^3/- /10^3	10^3/- /10^3

Table B.1 – Training details

Actor	Encoder	Decoder
Input 100 × 100 imgs F.C. output 51 × 51 Sigmoid	Input 51 × 51 imgs F.C. output 2× latent dim. no activ.	Input latent dim. vector F.C. output 51 × 51 Sigmoid F.C. output 100 × 100 Sigmoid

Table B.2 – Fully-variational network architecture

$\text{Actor}_{\text{real/imag}}$	$\text{Model}_{\text{real/imag}}$
Input 200 × 200 imgs F.C. output 51 × 51 Sigmoid	Input 51 × 51 imgs F.C. output 200 × 200 Sigmoid

Table B.3 – Maximum likelihood network architecture

C Appendix for chapter 4

C.1 Information-bottleneck formalism

To obtain a bound on the IB objective, we use the Markov chain constraint and the factorized joint distribution:

$$p(\mathbf{X},\mathbf{Y},\mathbf{Z}) = p(\mathbf{Y}|\mathbf{X},\mathbf{Z})\,p(\mathbf{Z}|\mathbf{X})\,p(\mathbf{X}) = p(\mathbf{Y}|\mathbf{X})\,p(\mathbf{Z}|\mathbf{X})\,p(\mathbf{X}) \tag{C.1}$$

to expand the mutual information terms in $\mathscr{L}_{IB} = \max\big[\, I(\mathbf{Z},\mathbf{Y}) - \beta I(\mathbf{Z},\mathbf{X})\big]$. Henceforth, we use the stochastic encoder $p_\phi(\mathbf{Z}|\mathbf{X})$ parameterized by ϕ as an approximation for $p(\mathbf{Z}|\mathbf{X})$. Starting with $I(\mathbf{Z},\mathbf{X})$, we have:

$$\begin{aligned} I(\mathbf{Z},\mathbf{X}) &= \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z})\; \log \frac{p(\mathbf{X},\mathbf{Z})}{p(\mathbf{X})\,p(\mathbf{Z})} \\ &= \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}) - \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X}|\mathbf{Z})\,p(\mathbf{Z}) \log p(\mathbf{Z}) \\ &= \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}) - \int d\mathbf{Z}\; p(\mathbf{Z}) \log p(\mathbf{Z}) \end{aligned} \tag{C.2}$$

where the second term on the RHS of Eq. C.2 is the entropy $H(\mathbf{Z})$. In practice computation of $H(\mathbf{Z})$ might be intractable (even though $P(\mathbf{Z})$ is well defined). Therefore, a variational approximation $\rho(\mathbf{Z})$ is used in place of $p(\mathbf{Z})$ such that $\mathrm{KL}(p(\mathbf{Z}),\rho(\mathbf{Z}))$ is minimal. Therefore,

with $\mathrm{KL}(p(\mathbf{Z}), \rho(\mathbf{Z})) \geq 0$, we have:

$$\begin{aligned} I(\mathbf{Z},\mathbf{X}) &= \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}) - \int d\mathbf{Z}\; p(\mathbf{Z}) \log p(\mathbf{Z}) \\ &\leq \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X},\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{X}) - \int d\mathbf{Z}\; p(\mathbf{Z}) \log \rho(\mathbf{Z}) \\ &= \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{X}|\mathbf{Z})\, p(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{\rho(\mathbf{Z})} = \int d\mathbf{X}\, d\mathbf{Z}\; p(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{\rho(\mathbf{Z})} \,. \end{aligned} \tag{C.3}$$

Using the stochastic encoder $p_\phi(\mathbf{Z}|\mathbf{X})$, an upper bound on $I(\mathbf{Z},\mathbf{X})$ reads as:

$$I(\mathbf{Z},\mathbf{X}) \leq \int d\mathbf{X}\, d\mathbf{Z}\; p_\phi(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X}) \log \frac{p_\phi(\mathbf{Z}|\mathbf{X})}{\rho(\mathbf{Z})} \,. \tag{C.4}$$

Moving on to the term $I(\mathbf{Z},\mathbf{Y})$, we have:

$$\begin{aligned} I(\mathbf{Z},\mathbf{Y}) &= \int d\mathbf{Y}\, d\mathbf{Z}\; p(\mathbf{Y},\mathbf{Z})\; \log \frac{p(\mathbf{Y},\mathbf{Z})}{p(\mathbf{Y})\, p(\mathbf{Z})} \\ &= \int d\mathbf{Y}\, d\mathbf{Z}\; p(\mathbf{Y},\mathbf{Z}) \log p(\mathbf{Y}|\mathbf{Z}) - \int d\mathbf{Y}\; p(\mathbf{Y}) \log p(\mathbf{Y}) \end{aligned} \tag{C.5}$$

where the second term on the RHS of Eq. C.5 is the entropy $H(\mathbf{Y})$. In practice computation of $p(\mathbf{Y},\mathbf{Z})$ and $p(\mathbf{Y}|\mathbf{Z})$ might be intractable (even though they are well defined). From Eq. C.1, $p(\mathbf{Y},\mathbf{Z})$ is written as $p(\mathbf{Y},\mathbf{Z}) = \int d\mathbf{X}\; p(\mathbf{Y}|\mathbf{X})\, p_\phi(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X})$. Additionally, a variational approximation $q_\theta(\mathbf{Y}|\mathbf{Z})$ is used in place of $p(\mathbf{Y}|\mathbf{Z})$ such that $\mathrm{KL}(q_\theta(\mathbf{Y}|\mathbf{Z}), p(\mathbf{Y}|\mathbf{Z})$ is minimal. Therefore, with $\mathrm{KL}(q_\theta(\mathbf{Y}|\mathbf{Z}), p(\mathbf{Y}|\mathbf{Z})) \geq 0$, we have:

$$\begin{aligned} I(\mathbf{Z},\mathbf{Y}) &= \int d\mathbf{Y}\, d\mathbf{Z}\; p(\mathbf{Y},\mathbf{Z}) \log p(\mathbf{Y}|\mathbf{Z}) + H(\mathbf{Y}) \\ &\geq \int d\mathbf{Y}\, d\mathbf{Z}\, d\mathbf{X}\; p(\mathbf{Y}|\mathbf{X})\, p_\phi(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X}) \log q_\theta(\mathbf{Y}|\mathbf{Z}) + H(\mathbf{Y}) \end{aligned} \tag{C.6}$$

With the bounds on $I(\mathbf{Z},\mathbf{Y})$ and $I(\mathbf{Z},\mathbf{X})$, the IB objective reads as:

$$\mathscr{L}_{IB} = \max\big[\, I(\mathbf{Z},\mathbf{Y}) - \beta I(\mathbf{Z},\mathbf{X}) \big] \geq \max_{\theta,\phi} \big[\, I_{IB} \big]$$

where

$$I_{IB} = \int d\mathbf{Y}\, d\mathbf{Z}\, d\mathbf{X}\, p(\mathbf{Y}|\mathbf{X})\, p_\phi(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X}) \log q_\theta(\mathbf{Y}|\mathbf{Z}) - \beta \int d\mathbf{X}\, d\mathbf{Z}\, p_\phi(\mathbf{Z}|\mathbf{X})\, p(\mathbf{X}) \log \frac{p_\phi(\mathbf{Z}|\mathbf{X})}{\rho(\mathbf{Z})} \tag{C.7}$$

As explained in the main text, we assume the joint distribution $p(\mathbf{X}, \mathbf{Y})$ is approximated by:

$$p(\mathbf{X}, \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^{T} \delta(\mathbf{X} - g(x_{1:\tau(t)}))\delta(\mathbf{Y} - y_t), \tag{C.8}$$

which allows the lower bound I_{IB} to be approximated by:

$$\begin{aligned} I_{IB} &\approx \frac{1}{T} \sum_{t=1}^{T} \Big[\int d\mathbf{Z}\, p_\phi(\mathbf{Z}|g(x_{1:\tau(t)})) \log q_\theta(y_t|\mathbf{Z}) - \beta \int d\mathbf{Z}\, p_\phi(\mathbf{Z}|g(x_{1:\tau(t)})) \log \frac{p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))}{\rho(\mathbf{Z})} \Big] \\ &= \frac{1}{T} \sum_{t=1}^{T} \Big[\mathbb{E}_{p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))} [\log q_\theta(y_t|\mathbf{Z})] - \beta D_{KL}[p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))||\rho(\mathbf{Z})] \Big] \end{aligned} \tag{C.9}$$

Finally, we enforce the Gaussian Process (GP) prior to derive the IB lower bound:

$$I_{IB} \approx \frac{1}{T} \sum_{t=1}^{T} \Big[\mathbb{E}_{p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))} [\log q_\theta(y_t|\mathbf{Z})] - \beta D_{KL}[p_\phi(\mathbf{Z}|g(x_{1:\tau(t)}))||\mathcal{GP}_{\mathbf{Z}}(\mathbf{0}, \boldsymbol{\Sigma})] \Big]. \tag{C.10}$$

$Z^l \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \in \mathbb{R}^{\mathcal{T}}$

C.2 Network architecture and optimization

The hyper parameters of the forward and backward networks, optimizers as well as training epochs used for training is summarized in Table C.1. Architecture of the networks is presented in Table C.2.

	Retina stimuli optimization
Optimizer	Adam
Learning rate	10^{-4}
VAE's β	500/450
Latent space dim.	100
Train/val/test batch size	20/- /10^3
Train/val/test batch num.	10^3/- /1

Table C.1 – Training details

Actor	Encoder	Decoder
Input 50 × 50 × 1000 seq. of imgs	Input 50 × 50 × 1000 seq. of imgs	F.C. output 16 × 12 × 12 Relu
3 × 3 conv. 64 s. 1 same Relu	3 × 3 conv. 64 s. 1 same Relu	3 × 3 conv. 32 s. 1 same Relu
2 × 2 maxpool	2 × 2 maxpool	2 × 2 Upsampling
3 × 3 conv. 32 s. 1 same Relu	3 × 3 conv. 32 s. 1 same Relu	4 sided zero pad.
2 × 2 maxpool	2 × 2 maxpool	3 × 3 conv. 64 s. 1 same Relu
3 × 3 conv. 16 s. 1 same Relu	3 × 3 conv. 16 s. 1 same Relu	2 × 2 Upsampling
F.C. output Bottleneck(1/4/9) No activ.	F.C. output 2× Latent dim. No activ.	3 × 3 conv. 1 s. 1 same Sigmoid
F.C. output 16 × 12 × 12		21 × 21 conv. 4 s. 1 no pad. no activ
3 × 3 conv. 32 s. 1 same Relu		40 × 1 1D-conv. 4 s. 1 same Relu
2 × 2 Upsampling		15 × 15 conv. 4 s. 1 no pad. Relu
4 sided zero pad.		F.C. output 9 Exponential activ.
3 × 3 conv. 64 s. 1 same Relu		
2 × 2 Upsampling		
3 × 3 conv. 1 s. 1 same Sigmoid		

Table C.2 – Retina network architecture

Bibliography

[1] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 1437–1443, 2017.

[2] Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light: Science & Applications*, vol. 7, no. 2, pp. 17 141–17 141, 2018.

[3] A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," *Optica*, vol. 4, no. 9, pp. 1117–1125, 2017.

[4] N. Brackbill, C. Rhoades, A. Kling, N. P. Shah, A. Sher, A. M. Litke, and E. Chichilnisky, "Reconstruction of natural images from responses of primate retinal ganglion cells," *Elife*, vol. 9, e58516, 2020.

[5] O. Jeromin, M. S. Pattichis, and V. D. Calhoun, "Optimal compressed sensing reconstructions of fmri using 2d deterministic and stochastic sampling geometries," *Biomedical engineering online*, vol. 11, no. 1, pp. 1–36, 2012.

[6] Y. Li, Y. Xue, and L. Tian, "Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media," *Optica*, vol. 5, no. 10, pp. 1181–1190, 2018.

[7] Y. Li, S. Cheng, Y. Xue, and L. Tian, "Displacement-agnostic coherent imaging through scatter with an interpretable deep neural network," *Optics Express*, vol. 29, no. 2, pp. 2244–2257, 2021.

[8] N. Thanh, Y. Xue, Y. Li, L. Tian, and G. Nehmetallah, "Deep learning approach to fourier ptychographic microscopy," *Optics express*, 2018.

[9] Y. Xue, S. Cheng, Y. Li, and L. Tian, "Reliable deep-learning-based phase imaging with uncertainty quantification," *Optica*, vol. 6, no. 5, pp. 618–629, 2019.

[10] S. Li, M. Deng, J. Lee, A. Sinha, and G. Barbastathis, "Imaging through glass diffusers using densely connected convolutional networks," *Optica*, vol. 5, no. 7, pp. 803–813, 2018.

[11] A. Goy, K. Arthur, S. Li, and G. Barbastathis, "Low photon count phase retrieval using deep learning," *Physical review letters*, vol. 121, no. 24, p. 243 902, 2018.

[12] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," *Nature methods*, vol. 16, no. 1, pp. 103–110, 2019.

[13] G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," *Optica*, vol. 6, no. 8, pp. 921–943, 2019.

[14] L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, and S. Baccus, "Deep learning models of the retinal response to natural scenes," *Advances in neural information processing systems*, vol. 29, pp. 1369–1377, 2016.

[15] B. Rahmani, D. Loterie, G. Konstantinou, D. Psaltis, and C. Moser, "Multimode optical fiber transmission with a deep learning network," *Light: Science & Applications*, vol. 7, no. 1, pp. 1–11, 2018.

[16] E. Spitz and A. Werts, "Transmission des images à travers une fibre optique," *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences Serie B*, vol. 264, no. 14, pp. 1015–+, 1967.

[17] A. Yariv, "On transmission and recovery of three-dimensional image information in optical waveguides," *JOSA*, vol. 66, no. 4, pp. 301–306, 1976.

[18] A. Gover, C. Lee, and A. Yariv, "Direct transmission of pictorial information in multimode optical fibers," *JOSA*, vol. 66, no. 4, pp. 306–311, 1976.

[19] G. J. Dunning and R. Lind, "Demonstration of image transmission through fibers by optical phase conjugation," *Optics letters*, vol. 7, no. 11, pp. 558–560, 1982.

[20] A. Friesem, U. Levy, and Y. Silberberg, "Parallel transmission of images through single optical fibers," *Proceedings of the IEEE*, vol. 71, no. 2, pp. 208–221, 1983.

[21] I. N. Papadopoulos, S. Farahi, C. Moser, and D. Psaltis, "Focusing and scanning light through a multimode optical fiber using digital phase conjugation," *Optics express*, vol. 20, no. 10, pp. 10 583–10 590, 2012.

[22] ——, "High-resolution, lensless endoscope based on digital scanning through a multimode optical fiber," *Biomedical optics express*, vol. 4, no. 2, pp. 260–270, 2013.

[23] Y. Choi, C. Yoon, M. Kim, T. D. Yang, C. Fang-Yen, R. R. Dasari, K. J. Lee, and W. Choi, "Scanner-free and wide-field endoscopic imaging by using a single multimode optical fiber," *Physical review letters*, vol. 109, no. 20, p. 203 901, 2012.

[24] A. M. Caravaca-Aguirre, E. Niv, D. B. Conkey, and R. Piestun, "Real-time resilient focusing through a bending multimode fiber," *Optics express*, vol. 21, no. 10, pp. 12 881–12 887, 2013.

[25] R. Y. Gu, R. N. Mahalati, and J. M. Kahn, "Design of flexible multi-mode fiber endoscope," *Optics express*, vol. 23, no. 21, pp. 26 905–26 918, 2015.

[26] D. Loterie, S. Farahi, I. Papadopoulos, A. Goy, D. Psaltis, and C. Moser, "Digital confocal microscopy through a multimode fiber," *Optics express*, vol. 23, no. 18, pp. 23 845–23 858, 2015.

[27] R. Di Leonardo and S. Bianchi, "Hologram transmission through multi-mode optical fibers," *Optics express*, vol. 19, no. 1, pp. 247–254, 2011.

[28] T. Čižmár and K. Dholakia, "Shaping the light transmission through a multimode optical fibre: complex transformation analysis and applications in biophotonics," *Optics express*, vol. 19, no. 20, pp. 18 871–18 884, 2011.

[29] ——, "Exploiting multimode waveguides for pure fibre-based imaging," *Nature communications*, vol. 3, no. 1, pp. 1–9, 2012.

[30] S. Bianchi and R. Di Leonardo, "A multi-mode fiber probe for holographic micromanipulation and microscopy," *Lab on a Chip*, vol. 12, no. 3, pp. 635–639, 2012.

[31] E. R. Andresen, G. Bouwmans, S. Monneret, and H. Rigneault, "Toward endoscopes with no distal optics: video-rate scanning microscopy through a fiber bundle," *Optics letters*, vol. 38, no. 5, pp. 609–611, 2013.

[32] S. Popoff, G. Lerosey, M. Fink, A. C. Boccara, and S. Gigan, "Image transmission through an opaque material," *Nature communications*, vol. 1, no. 1, pp. 1–5, 2010.

[33] M. N'Gom, M.-B. Lien, N. M. Estakhri, T. B. Norris, E. Michielssen, and R. R. Nadakuditi, "Controlling light transmission through highly scattering media using semi-definite programming as a phase retrieval computation method," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[34] M. N'Gom, T. B. Norris, E. Michielssen, and R. R. Nadakuditi, "Mode control in a multimode fiber through acquiring its transmission matrix from a reference-less optical system," *Optics letters*, vol. 43, no. 3, pp. 419–422, 2018.

[35] Z. Zhang, Z. You, and D. Chu, "Fundamentals of phase-only liquid crystal on silicon (lcos) devices," *Light: Science & Applications*, vol. 3, no. 10, e213–e213, 2014.

[36] E. Van Putten, I. M. Vellekoop, and A. Mosk, "Spatial amplitude and phase modulation using commercial twisted nematic lcds," *Applied optics*, vol. 47, no. 12, pp. 2076–2081, 2008.

[37] M. Mounaix and J. Carpenter, "Control of the temporal and polarization response of a multimode fiber," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[38] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2017, pp. 2921–2926.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[40] N. Borhani, E. Kakkava, C. Moser, and D. Psaltis, "Learning to see through multimode fibers," *Optica*, vol. 5, no. 8, pp. 960–966, 2018.

[41] E. Kakkava, B. Rahmani, N. Borhani, U. Teğin, D. Loterie, G. Konstantinou, C. Moser, and D. Psaltis, "Imaging through multimode fibers using deep learning: the effects of intensity versus holographic recording of the speckle pattern," *Optical Fiber Technology*, vol. 52, p. 101 985, 2019.

[42] P. Caramazza, O. Moran, R. Murray-Smith, and D. Faccio, "Transmission of natural scene images through a multimode fibre," *Nature communications*, vol. 10, no. 1, pp. 1–6, 2019.

[43] C. Zhu, E. A. Chan, Y. Wang, W. Peng, R. Guo, B. Zhang, C. Soci, and Y. Chong, "Image reconstruction through a multimode fiber with a simple neural network architecture," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[44] P. Fan, T. Zhao, and L. Su, "Deep learning the high variability and randomness inside multimode fibers," *Optics express*, vol. 27, no. 15, pp. 20 241–20 258, 2019.

[45] S. Resisi, S. M. Popoff, and Y. Bromberg, "Image transmission through a flexible multimode fiber by deep learning," *arXiv preprint arXiv:2011.05144*, 2020.

[46] J. Zhao, X. Ji, M. Zhang, X. Wang, Z. Chen, Y. Zhang, and J. Pu, "High-fidelity imaging through multimode fibers via deep learning," *Journal of Physics: Photonics*, vol. 3, no. 1, p. 015 003, 2021.

[47] E. Kakkava, N. Borhani, B. Rahmani, U. Teğin, C. Moser, and D. Psaltis, "Deep learning-based image classification through a multimode fiber in the presence of wavelength drift," *Applied Sciences*, vol. 10, no. 11, p. 3816, 2020.

[48] ——, "Wavelength independent image classification through a multimode fiber using deep neural networks," in *The European Conference on Lasers and Electro-Optics*, Optical Society of America, 2019, ci_2_1.

[49] E. Kakkava, N. Borhani, B. Rahmani, U. Tegin, C. Moser, and D. Psaltis, "Efficient image classification through a multimode fiber using deep neural networks in presence of wavelength drifting," in *Computational Optical Sensing and Imaging*, Optical Society of America, 2019, CW1A–4.

[50] Y. Luo, S. Yan, H. Li, P. Lai, and Y. Zheng, "Towards smart optical focusing: deep learning-empowered dynamic wavefront shaping through nonstationary scattering media," *Photonics Research*, vol. 9, no. 8, B262–B278, 2021.

[51] M. Wei, G. Tang, J. Liu, L. Zhu, J. Liu, C. Huang, J. Zhang, L. Shen, and S. Yu, "Neural network based perturbation-location fiber specklegram sensing system towards applications with limited number of training samples," *Journal of Lightwave Technology*, vol. 39, no. 19, pp. 6315–6326, 2021.

[52] Y. Liu, G. Li, Q. Qin, Z. Tan, M. Wang, and F. Yan, "Bending recognition based on the analysis of fiber specklegrams using deep learning," *Optics & Laser Technology*, vol. 131, p. 106 424, 2020.

[53] T. Würfl, F. C. Ghesu, V. Christlein, and A. Maier, "Deep learning computed tomography," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 432–440.

[54] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[55] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[56] A. Mousavi and R. G. Baraniuk, "Learning to invert: signal recovery via deep convolutional networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 2272–2276.

[57] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," *arXiv preprint arXiv:1703.03208*, 2017.

[58] V. Shah and C. Hegde, "Solving linear inverse problems using gan priors: an algorithm with provable guarantees," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4609–4613.

[59] D. Van Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *arXiv preprint arXiv:1806.06438*, 2018.

[60] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[61] R. Heckel and P. Hand, "Deep decoder: concise image representations from untrained non-convolutional networks," *arXiv preprint arXiv:1810.03982*, 2018.

[62] B. Rahmani, D. Psaltis, and C. Moser, "Variational framework for partially-measured physical system control," *4th workshop of Machine Learning for Physical Sciences NeurIPS*, 2021.

[63] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-vae: learning basic visual concepts with a constrained variational framework," 2016.

[64] H. C. Carver, A. O'TOOLE, and T. RAIFORD, *The annals of mathematical statistics*. Edwards Bros., 1930.

[65] N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. Chichilnisky, and L. Paninski, "Neural networks for efficient bayesian decoding of natural images from retinal neurons," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6434–6445, 2017.

[66] P. Bashivan, K. Kar, and J. J. DiCarlo, "Neural population control via deep image synthesis," *Science*, vol. 364, no. 6439, 2019.

[67] B. Rahmani, D. Loterie, E. Kakkava, N. Borhani, U. Teğin, D. Psaltis, and C. Moser, "Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 403–410, 2020.

[68] A. Wilson and A. Di Polo, "Gene therapy for retinal ganglion cell neuroprotection in glaucoma," *Gene therapy*, vol. 19, no. 2, pp. 127–136, 2012.

[69] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[70] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[71] R. Bamler and S. Mandt, "Structured black box variational inference for latent time series models," *arXiv preprint arXiv:1707.01069*, 2017.

[72] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.

[73] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.

[74] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[75] E. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Network: computation in neural systems*, vol. 12, no. 2, p. 199, 2001.

[76] J. W. Pillow, L. Paninski, V. J. Uzzell, E. P. Simoncelli, and E. Chichilnisky, "Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model," *Journal of Neuroscience*, vol. 25, no. 47, pp. 11 003–11 013, 2005.

[77] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[78] N. Maheswaranathan, L. T. McIntosh, H. Tanaka, S. Grant, D. B. Kastner, J. B. Melander, A. Nayebi, L. Brezovec, J. Wang, S. Ganguli, *et al.*, "The dynamic neural code of the retina for natural scenes," *BioRxiv*, p. 340 943, 2019.

[79] E. Batty, J. Merel, N. Brackbill, A. Heitman, A. Sher, A. Litke, E. Chichilnisky, and L. Paninski, "Multilayer recurrent network models of primate retinal ganglion cell responses," 2016.

[80] N. Shah, S. Madugula, P. Hottowy, A. Sher, A. Litke, L. Paninski, and E. Chichilnisky, "Efficient characterization of electrically evoked responses for neural interfaces," *Advances in Neural Information Processing Systems*, vol. 32, pp. 14 444–14 458, 2019.

[81] E. Y. Walker, F. H. Sinz, E. Cobos, T. Muhammad, E. Froudarakis, P. G. Fahey, A. S. Ecker, J. Reimer, X. Pitkow, and A. S. Tolias, "Inception loops discover what excites neurons most using deep predictive models," *Nature neuroscience*, vol. 22, no. 12, pp. 2060–2065, 2019.

[82] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[83] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, PMLR, 2015, pp. 448–456.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[85] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Babak Rahmani

EPFL STI IMT LAPD BM 4110 Station 17
CH-1015 Lausanne Switzerland
+41 78 752 9254
babak.rahmani@epfl.ch
Github *Linkedin* *Webpage*

Education

2018–now: **PhD, Electrical Engineering**, *EPFL*, Lausanne, Switzerland.
- Inverse Problems in Computational Imaging, Applied Machine Learning and Deep Neural Networks.
- Credits obtained: **26**.

2014–2016: **Master of Electrical Engineering**, *Sharif University of Technology*, Tehran, Iran.
- M.Sc. in Engineering with a focus on Microwave and Optical Communications.
- GPA: **17.77/20.00 (3.79/4)**.

2010–2014: **Bachelor of Electrical Engineering**, *Tehran University*, Tehran, Iran.
- B.Sc. in Engineering with a focus on Telecommunications.
- GPA: **18.03/20.00 (3.88/4)**.
- Thesis: Device to Device Communication.

Selected Ph.D. Courses

Statistical physics for optimization and learning .....5.75/6.00.....Instructor: Profs. Florent Krzakala and Lenka Zdeborová.

Adaptation and Learning.....Passed.....Instructor: Prof. Ali H. Sayed.

Advance Topics in Machine Learning .....5.25/6.00.....Instructor: Prof. Pascal Frossard, Cevher Volkan, and others.

Machine Learning for Engineers.....5.75/6.00.....Instructor: Prof. François Fleuret and others.

Theory and Methods for Reinforcement Learning.....5.50/6.00.....Instructor: Prof. Cevher Volkan.

Deep Learning For Natural Language Processing.....Ongoing.....Instructor: Dr. James Henderson.
Neural Computation.....Audited.....Instructor: Prof. Bruno Olshausen.

Publications* & Patents

* Please visit my *Google Scholar* for an updated version of the publications as well as conference proceedings.

2021 **B. Rahmani**, D. Psaltis, C. Moser, Variational framework for partially-measured physical system control: examples of vision neuroscience and optical random media, Workshop on Machine Learning and the Physical Sciences, **NeurIPS 2021**, Vancouver, 2021.

2021 **B. Rahmani**, D. Loterie, E. Kakkava, N. Borhani, U. Teğin, D. Psaltis, C. Moser, Partially-measured physical system characterization with neural networks, **Invited talk**, SPIE San Diego, 2021.

2021 **B. Rahmani**, D. Loterie, E. Kakkava, N. Borhani, U. Teğin, D. Psaltis, C. Moser, Multimode fiber projector with neural networks, **Conference presentation**, SPIE San Francisco, 2021.

2020 **B. Rahmani**, D. Loterie, E. Kakkava, N. Borhani, U. Teğin, D. Psaltis, C. Moser, Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media, **Nature Machine Intelligence**, 2(7), 2020.

2020 **B. Rahmani**, D. Loterie, E. Kakkava, N. Borhani, U. Teğin, D. Psaltis, C. Moser, Multimode fiber projection with machine learning, **Conference presentation**, OSA Vancouver, 2020.

2020 U. Teğin, **B. Rahmani**, E. Kakkava, N. Borhani, D. Psaltis, C. Moser, Controlling spatiotemporal nonlinearities in multimode fibers with deep neural networks, **APL Photonics**, 5(3), 2020.

2020 E. Kakkava, **B. Rahmani**, N. Borhani, U. Teğin, C. Moser, D. Psaltis, Deep Learning-Based Image Classification through a Multimode Fiber in the Presence of Wavelength Drift, **Applied Sciences**, 10(11), 2020.

2020 O. Hemmatyar, M. Abbassi, **B. Rahmani**, M. Memarian, K. Mehrany, Wide-band/angle blazed dual mode metallic groove gratings, **IEEE Transactions on Antennas and Propagation**, 2020.

2019 E. Kakkava, **B. Rahmani**, N. Borhani, U. Teğin, D. Loterie, G. Konstantinou, C. Moser, D. Psaltis, Imaging through multimode fibers using deep learning: The effects of intensity versus holographic recording of the speckle pattern, **Optical Fiber Technology**, 52, 2019.

2019 U. Teğin, **B. Rahmani**, E. Kakkava, D. Psaltis, C. Moser, Spatiotemporal self-similar fiber laser, **Optica**, 6(11), 2019.

2019 M. Tavakol, **B. Rahmani**, A. Khavasi, Terahertz quarter wave-plate metasurface polarizer based on arrays of graphene ribbons, **IEEE Photonics Technology Letters**, 31(12), 2019.

2018 **Rahmani B.**, Loterie D., Konstantinou G., Psaltis D., Moser C., Multimode optical fiber transmission with a deep learning network, **Nature Light: Science & Applications**, 7(69), 2018.

2018 M. Tavakol, **B. Rahmani**, A. Khavasi, Tunable polarization converter based on one-dimensional graphene metasurfaces, **JOSA B**, 35(10), 2018.

2018 **Rahmani B**, K. Mehrany., Modeling of Periodic Array of Cut-through Slits with Sinusoidal Surface Conductivity at the Interfaces of an Anisotropic Medium, **IEEE Transactions on Antennas and Propagation**, 66(10), 2018.

2017 O. Hemmatyar, **B. Rahmani**, A. Bagheri, A. Khavasi, Phase Resonance Tuning and Multi-Band Absorption Via Graphene-Covered Compound Metallic Gratings, **IEEE Journal of Quantum Electronics**, 53(5), 2017.

2017 A. Bagheri, **B. Rahmani**, A. Khavasi, Effect of Graphene on the Absorption and Extraordinary Transmission of light in One Dimensional Metallic Gratings, **IEEE Journal of Quantum Electronics**, 53(3), 2017.

2016 **Rahmani B**, A. Bagheri., A. Khavasi., K. Mehrany, Effective Medium Theory for Graphene-covered Metallic Gratings, **Journal of Optics**, 18(10), 2016.

Patents

2019 C. Moser, **B. Rahmani**, D. Psaltis, System and method for projecting images through scattering media, Under evaluation.

Research Experience

EPFL, Lausanne, Switzerland

February 2021 – Ongoing ***Blind Deconvolution***.
As the research project of the Statistical physics for optimization and learning course, I am working on the blind deconvolution problem which involves recovering the signal under Gaussian noisy channel in the Bayesian setting using probabilistic graphical models, replica method and approximate message passing algorithms.

August 2020 – Ongoing ***RetinaAI***.
I am involved with a neuroscience-related project in which the goal is to use data-driven methods based on machine learning to control the spikings of Retina Ganglion Cells (RGCs) via intelligent stimulation of the photo-receptors so as to produce the same spiking of RGCs evoked by stimulation via natural images but with certain constraints. During the course of the project, I have been exposed to various concepts in machine learning such as Representation Learning and Variation Autoencoders. This ongoing project also required me to build the hardware (optical setup and stimulation/recording apparatus) needed for collecting data.

January 2019 – August 2021 ***Neural networks for control: nonlinear time-varying complex media control***.
Developing semi-supervised neural-network based controllers for online control of time-varying media of optical fibers. During the course of this project, I was exposed to several concepts in Deep Learning ranging from Auto-encoders, adverserial training and untrained neural networks to dealing with the nuisances of big data from real-world systems.**Results published in Nature Machine Intelligence.**

January 2018 – August 2018 ***Neural networks for inference: computational image reconstruction***.
Developing state-of-the-art deep neural networks for image reconstruction in the multimodal complex media of optical fibers. During the course of this project, I was exposed to concepts such as super-resolution imaging and denoising with neural networks. **Results published in Nature Light science and applications.**

Master's projects

Sharif University of Technology, Tehran, Iran

Sep. 2014 – Dec 2017 ***Integrated Photonics Lab (IPL)***.
Research Assistant, EE Department, Sharif University of Technology, Tehran, Iran. Design, simulation and analysis of Metallic Gratings and Graphene-based Structures.

Advisors : **Drs. Khashayar Mehrany, Mohammad Memarian, Amin Khavasi**, *Department of Electrical Engineering*, Sharif University of Technology.

University of Tehran, Tehran, Iran

May 2013 – Oct. 2013 ***Type Approval Antenna Lab***.
Intern Assistant, ECE Department, University of Tehran, Tehran, Iran. Working on Antenna pattern measurements such as: Scalar Pattern Measurement including Amplitude and Polarization pattern, Directivity, and Gain measurements Response Measurement including Antenna VSWR, and Material measurements CTIA Measurement including OTA-TRP (EIRP).

Honors & Awards

2017 **Cornell University Fellowships Award** for Ph.D. program in Electrical Engineering, USA.

2017 **Graduate Assistantship Award** for Ph.D. program in Electrical Engineering at Georgia Institute of Technology, USA.

2014 **Scholarship** for M.Sc. in the Communications major at University of Tehran, Tehran, Iran. Entrance examination waived as an award for being among the **Top-10% students (out of 120+)**.

2010 Ranked **187th among approximately 150,000 participants** in the nationwide university entrance examination in Mathematics and Physics fields for B.Sc. degree.

2007 Admitted in the first stage of nationwide Mathematics Olympiad for High school students in Iran.

Academic Achievements & Recognitions

2020 **News coverage** of First author paper "Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media.

2020 First author paper "Multimode optical fiber transmission with a deep learning network" was recognized as one of the **top downloaded papers in top-tier Nature journal Light: Science & Applications in 2019**.

2019 **First prize** for EPFL's Electrical Engineering department (EDEE) end-of-the-year poster competition.

2019 **Invited talk** on the use of **Deep Learning for solving inverse problem and computational imaging** at a major Photonics conference venue, SPIE Photonics West 2019, San Francisco, California, USA.

2018 **Second prize** for EPFL's Electrical Engineering department (EDEE) end-of-the-year poster competition.

Leadership & Voluntary Experiences

2020-ongoing PhD Student Representative of the EPFL's Electrical Engineering Students.
Reviewer of various journals/venues such as Nature, IEEE, APL, OSA.

Familiarity with Computer Systems and CAD Software

Machine Learning	Tensorflow (pro), Pytorch (limited), scikit-learn
Programming Languages	Python, C, C++, MATLAB, Assembly, Verilog and FPGA programming
Windows Software	Microsoft Office Package, AutoCAD
Operating System	Microsoft Windows, Linux

Teaching Assistantship*

Spring, 2018 : **Teaching assistant of Math and Physics**, EPFL.
Spring, 2015 : **Teaching assistant of Engineering Mathematics**, Sharif University of Technology.
fall, 2015 : **Teaching assistant of Engineering Mathematics**, Sharif University of Technology.
Spring, 2013 : **Teaching assistant of Engineering Mathematics**, Tehran University.
Spring, 2013 : **Teaching assistant of Electrical Circuit II**, Tehran University.
Fall, 2012 : **Project designer and lab teaching assistant of Electronics I**, Tehran University.

* Held tutorial session, assisted students with laboratory experiments, marked assignments and exams.

Referees

Prof. Christophe Moser
Associate Professor, Department of Electrical Engineering
EPFL, Lausanne, Switzerland
✉ christophe.moser@epfl.ch

Prof. Demetri Psaltis
Full Professor, Department of Electrical Engineering
EPFL, Lausanne, Switzerland
✉ demetri.psaltis@epfl.ch

Prof. Khashayar Mehrany
Full Professor, Department of Electrical Engineering
Sharif University of Technology
✉ kh.mehrany@gmail.com