# Unwrapping Non-Locality in the Image Transmission Through Turbid Media

MOHAMMADRAHIM KAZEMZADEH,[1, 2] LIAM COLLARD,[1,2,3] FILIPPO PISANO,[1,4] LINDA PISCOPO,[1,5] CRISTIAN CIRACI, [1] MASSIMO DE VITTORIO, [1, 3, 5, 6,*] AND FERRUCCIO PISANELLO [1, 3, 6,*]

[1]*Istituto Italiano di Tecnologia Center for Biomolecular Nanotechnologies*
[2]*These two authors contributed equally*
[3]*RAISE Ecosystem, Genova, Italy*
[4]*Department of Physics and Astronomy "G.Galilei", University of Padua, Via F. Marzolo, 8, 35131 Padua (Italy)*
[5]*Dipartimento di Ingegneria dell'Innovazione, Università del Salento*
[6]*These two authors are co-last authors*
[*]*ferruccio.pisanello@iit.it*

**Abstract:** Achieving high-fidelity image transmission through turbid media is a significant challenge facing both the AI and photonic/optical communities. While this capability holds promise for a variety of applications, including data transfer, neural endoscopy, and multi-mode optical fiber-based imaging, conventional deep learning methods struggle to capture the nuances of light propagation, leading to weak generalization and limited reconstruction performance. To address this limitation, we investigated the non-locality present in the reconstructed images and discovered that conventional deep learning methods rely on specific features extracted from the training dataset rather than meticulously reconstructing each pixel. This suggests that they fail to effectively capture long-range dependencies between pixels, which are crucial for accurate image reconstruction. Inspired by the physics of light propagation in turbid media, we developed a global attention mechanism to approach this problem from a broader perspective. Our network harnesses information redundancy generated by peculiar non-local features across the input and output fiber facets. This mechanism enables a two-order-of-magnitude performance boost and high fidelity to the data context, ensuring accurate representation of intricate details in a pixel-to-pixel reconstruction rather than mere loss minimization.

## 1. Introduction

In the realm of multimodal waveguides, understanding the complex relationship between the input and the output electromagnetic fields is a long standing aim of the scientific community. For multimode optical fibers, the last decade has seen a strong research effort to solve this problem by interferomic techniques , with the main aim of widening their application to biomedical endoscopy and image transmission fields [1–8]. Coherent light injected into a MMF is coupled to multiple guided modes, which propagate with different phase delays, interfering in an apparently random fashion. At any arbitrary cross-section of the waveguide sufficiently far from the input, the distribution of the electric and magnetic fields does not resemble the distribution of the source in the excitation plane, and it takes the shape of a complex speckle pattern at the output facet. This happens even when a diffraction-limited point source (e.g. a focused laser beam) is injected in a specific point of the input facet, generating a speckle pattern on the entire output facet. This generates what can be referred at as a *non-local* correlation between the fields on these two optical planes, since a variation in a point of the input facet strongly affects the entire speckle pattern detected at the output, hindering the possibility of transmitting an image through a MMF.

A revolution in imaging transmission across these apparently random -strongly aberrating- media has been represented by the possibility to measure their transmission matrix through interferomic techniques [1–8]. This breakthrough has unlocked a host of fascinating applications

for MMFs, enabling far-field imaging [9, 10], holographic optical tweezers [11], and endoscopy [12, 13]. Remarkably, deep learning and machine learning methods have recently shown promise in leveraging multimodal optical fibers for image transmission, wavefront shaping, and holography at the distal end without any interferomic measurement. For instance, notable studies [14–20] demonstrated that a deep convolutional neural network based on a particular U-Net architecture [21] can be employed to reverse-engineer the distal end's speckle patterns back to their original generating phase pattern on a phase-only spatial light modulator (SLM) [14]. Moreover, the versatility of U-Net architecture has been employed in other related research endeavors, delving into factors influencing image transmission through the fiber, including bending [22, 23]. More recent research has explored the efficacy of state-of-the-art vision transformers (ViT) and its variants [24, 25], such as swine transformers [26, 27], in tackling the problem with improved effectiveness. Another study, [28] revealed that attention-based convolutional neural network architectures not only enhance the network's capacity to learn from smaller datasets but also improve its ability to generalize to unseen datasets through transfer learning. Additionally, multilayer perceptron architectures have been employed to address the transmission of natural scenes [16] through the fiber. Utilizing a single-layer neural network that incorporates complex-valued inputs, these architectures closely mimic the physics underlying wave propagation through the fiber. However, to circumvent the issue of exploding gradients and enable effective learning of the inverse transmission model, such networks require the implementation of additional regularizers. Although multilayer perceptrons are adept at capturing non-local patterns, they come with increased computational expenses in comparison to convolutional neural networks and vision transformers, especially in scenarios requiring deep networks. Therefore, all these deep-learning approaches do not efficiently capture the global features arising from the non-locality in the input-output relationship generated by light propagation in MMFs. The result is that currently available deep-learning methods rebuild the transmitted image on the base of specific features of the training dataset, hindering the possibility to obtain pixel-to-pixel reconstructions and effective transmission of grid-topology data.

Here, we provide evidence that the non-locality concept is accompanied by a redundancy in information transmission through turbid media, and that it can be exploited by deep neural networks to obtain high-fidelity reconstruction of transmitted images in a pixel-wise fashion. We showcase the redundancy principle describing that even a simple linear regression machine learning method, if aware of the non-locality, can be employed to reconstruct a transmitted phase image by probing only a small portion of the speckle patters generated at the fiber output. We then describe how non-locality alters the perception mechanism of classically employed deep neural network, showing that the resulting image reconstruction process is based on feature similarities rather than pixel-by-pixel reconstruction. To tackle this challenge we introduce a Global Awareness Module (GAM) that brings the notion of non-locality into complex CNN architectures, enabling pixel-to-pixel transmission and opens the way toward the application of similar topological non-localities in more general scattering problems.

## 2. Results

### *2.1. Non-locality in speckle generation through MMF*

To describe the non-locality in speckle generation through MMF we used the optical scheme depicted in Figure 1 (a more detailed version is presented in Supplementary Figure 1). The phase of continuous wave laser radiation is modulated by phase-only spatial light modulator (SLM), and the resulting wavefront coupled to a MMF (numerical aperture NA=0.22, core diameter 50$\mu$m) through an objective lens whose back focal plane is optically conjugated with the SLM screen, so that the Fourier transform of the pattern on SLM screen is projected in the MMF facet.

An alternative perspective on this coupling process is provided by ray optics, where each pixel of the SLM is uniquely inserted into the fiber at a specific angle with respect to the optical axis.

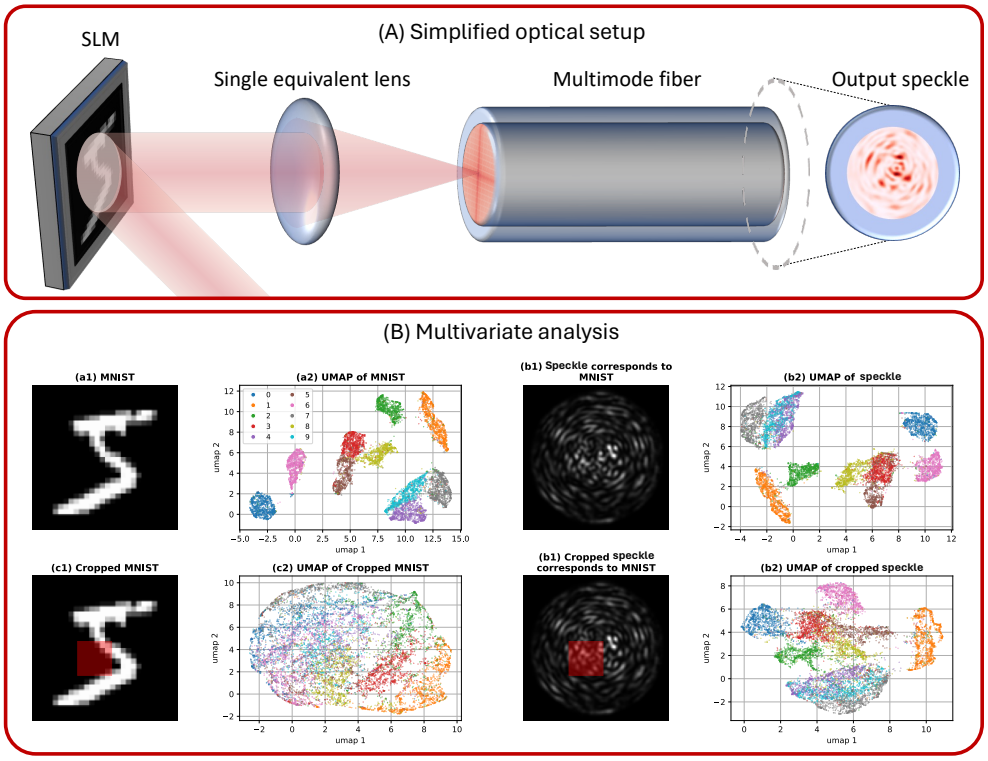


Fig. 1. Panel A: The simplified version of the optical setup employed in this study consists of a Spatial Light Modulator (SLM), a single lens forming a 2f system, and a multimode optical fiber. Panel B: (a1) and (b1) showcase an example of the MNIST digit dataset used as the SLM pattern and its corresponding speckle pattern, respectively. (a2) and (b2) display the UMAP projection of the datasets depicted in (a1) and (b1). (c1) and (d1) presented a random crop from the MNIST and speckle datasets, respectively, with the cropped section indicated using a red rectangle. (c2) and (d2) illustrate the UMAP projection of the cropped sections from (c1) and (d1), respectively.

Notably, this insertion must occur within the numerical aperture cone of the fiber to ensure the effective coupling of the SLM pixels. At the output of this system, a speckle pattern is obtained. Our goal is to establish a reliable pixel- to-pixel mapping of this speckle pattern back to its corresponding SLM pattern, exploiting the intrinsic non-locality relationships between the fields at these two planes.

To mathematically describe the electric field distribution at the distal end of the fiber as produced by the pixels of the SLM, we employ the following formulation:

$$\boldsymbol{E}(x,y) = \sum_{n=1}^{N} \boldsymbol{U}_n(x,y) e^{j\phi_n} \tag{1}$$

In this equation, $\boldsymbol{E}$ represented the total electric field vector at the distal end of the fiber, laying in $(x, y)$ plane (see Figure 1A for axes definition). The vector $\boldsymbol{U}$ corresponds to the electric field produced by the excitation from the n-th pixel in the SLM. Finally, $e$, $j$, and $\phi_n$, $N$, represent the Neper number, imaginary unit, the modulated phase corresponding to the n-th pixel in the SLM, and the total number of SLM's pixels, respectively. The relationship presented here stems from the linear time-invariant behavior of the governing Maxwell equations. However, it is essential to

acknowledge that the camera used in our experiments cannot directly capture the electric field itself; rather, it captures a quantity proportional to the electric field intensity. Consequently, our measured speckle pattern using the camera is expressed as follows:

$$|\boldsymbol{E}(x,y)|^2 = \sum_{m=1}^{N}\sum_{n=1}^{N}(\boldsymbol{U}_n(x,y).\boldsymbol{U}_m^*(x,y))e^{j(\phi_n-\phi_m)} \tag{2}$$

In this equation, the superscript * denotes the conjugated fields labeled with a subscript "m" for the m-th pixel in the SLM. Two crucial points warrant careful consideration. First, equations (1) and (2) reveal the non-local effect from the SLM screen to the speckle pattern. In this context, the n-th pixel of the SLM, though situated at a specific location on the SLM screen, can influence all locations in the generated speckle pattern. Second, the interference term $e^{j(\phi_n-\phi_m)}$ assumes significance. Let's first assume the phase difference to be small, i.e.,

$$(\phi_n - \phi_m) << 1 \tag{3}$$

now the equation (2) can be treated as a linear map between the phase pattern on the SLM and the resultant electric field. This observation prompted us to explore the efficacy of applying linear regression to establish a tool for image transmission systems through MMF. It is noteworthy that the existing literature has predominantly employed deep learning models to address this problem, given its inherently nonlinear and complex nature. To gain a clearer perspective on the validity of linear regression for this transformation, we can approximate the exponential function using the first two components of its Taylor expansion. This allows us to reformulate the aforementioned equation in the following manner:

$$\begin{bmatrix} |\boldsymbol{E}(x_1,y_1)|^2 \\ |\boldsymbol{E}(x_2,y_2)|^2 \\ \vdots \\ |\boldsymbol{E}(x_p,y_p)|^2 \end{bmatrix} = \begin{bmatrix} s_{1,1}(x_1,y_1) & s_{1,2}(x_1,y_1) & \cdots & s_{1,N}(x_1,y_1) \\ s_{2,1}(x_2,y_2) & s_{2,2}(x_2,y_2) & \cdots & s_{2,N}(x_2,y_2) \\ \vdots & \vdots & \ddots & \vdots \\ s_{p,1}(x_p,y_p) & s_{p,2}(x_p,y_p) & \cdots & s_{p,N}(x_p,y_p) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_p \end{bmatrix} \tag{4}$$

Here, the subscript $p$ pertains to the pixels captured by the camera in the speckle image. We sampled points from the $(x, y)$ space 'p' times and represented them in the form $(x_p,y_p)$, where the subscript '$p$' is indicative of the sampling process and does not denote diagonal selection from the $(x, y)$ matrix. Now, the real scalar value $s_{p,n}(x_p, y_p)$ can be calculated using equation 2, expressed as follows:

$$s_{p,n}(x_p,y_p) = j\sum_{m=1}^{N}(\boldsymbol{U}_n(x_p,y_p).\boldsymbol{U}_m^*(x_p,y_p) - \boldsymbol{U}_m(x_p,y_p).\boldsymbol{U}_n^*(x_p,y_p)) \tag{5}$$

Furthermore, the real value bias vector in Equation 3 can be denoted as:

$$B_p = \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}(\boldsymbol{U}_n(x_p,y_p).\boldsymbol{U}_m^*(x_p,y_p) + \boldsymbol{U}_m(x_p,y_p).\boldsymbol{U}_n^*(x_p,y_p)) \tag{6}$$

These equations highlight the presence of a linear relationship between the measured $|E(x_p, y_p)|^2$ and the phase within the Spatial Light Modulator (SLM). Consequently, the application of linear regression emerges as a viable approach for deducing the inverse transformation. It's crucial to underscore that this connection holds valid when the disparities in phase among SLM pixels remain relatively minor. Another significant insight derived from

Equations 3-5 is that if the number of pixels in the speckle image ($p$) surpasses the count of independently modulated phases ($N$), a definite linear correlation emerges among the elements in the electric field norm. Consequently, not all the p-th electric field norms detected by the camera catching the speckle patterns are required for reconstructing the modulated phase. This is a direct consequence of the intrinsic non-locality in the speckles generation.

This analytical observation can be further substantiated through experimental methods and data-driven techniques to showcase the general non-local behavior of the problem, depicted in Figure 1B. For this experiment, we propagated MNIST data through a 5 cm-long MMF fiber and captured the corresponding speckles generated at the distal end. Subsequently, we performed Uniform Manifold Approximation and Projection (UMAP), a form of unsupervised manifold learning, on both the MNIST dataset and its corresponding speckle data. As depicted in Figure 1B(top raw), it becomes evident that the data's underlying structure is retained through the fiber transmission.

However, a significant difference arises when we apply UMAP exclusively to a subset of both MNIST images and a similar portion of the speckle data, as shown in the bottom row of Figure 1B. In the UMAP projection of the MNIST dataset, the clusters representing each individual digit, as produced by the UMAP applied to the complete MNIST images, lose their integrity and collapse. This starkly contrasts with the UMAP projection of the cropped speckle data, where distinct clusters for each digit still persist. This phenomenon highlights redundancy in the speckle data and underscores its non-locality, as it encapsulates information from all pixels of the SLM across various sections of the speckle image. Supplementary Figure 2-4 highlight how this non-local correlation is distributed across the speckle pattern to the extent that the the ability to cluster the transmitted digit does not depend on the position of the employed speckles portion, and it still persists in the case of randomly selected points on the (x,y) plane.

### *2.2. Linear Regression*

In this section we describe to what extent the above-described non-locality can be exploited to obtain SLM image reconstruction with computational-efficient linear-regression, avoiding the use of complex deep learning models. To this aim, the speckle image was randomly sampled on a number of points equal to the count of independently modulated phases (N), being the minimum number that allows to reconstruct the phase of the SLM pixels based on equation (4). These points were utilized as inputs for the linear regression process, while their corresponding outputs were designated as the images that had generated the corresponding speckle patterns. This strategy of multivariate linear regression can be likened to the structure of a single-layer neural network, visually presented in Figure 2A(left). Subsequently, the reconstructed phase shifts at the output of the linear regression were organized into a two-dimensional array. This reshaping step was undertaken using our prior knowledge about the pixel locations on the SLM, resulting in the creation of comprehensible images. It's noteworthy that due to the inherent independence characterizing the reconstruction of each individual point, we were able to reconfigure the equations governing the reconstruction of each pixel. This reformulation was carried out in terms of the inner product of vectors, as depicted in Figure 2A(right). To train the linear regression we only employed 4000 images from each dataset and serve another 6000 as the testing set.

In Figure 2B, C, and D, we present the performance of linear regression in reconstructing three distinct datasets (MNIST, CIFAR, and a randomly generated dataset), each characterized by a unique phase distribution. Each linear regression model is individually trained on its respective dataset. For example, the linear regression model for MNIST is trained on 4000 MNIST data points and then tested on the remaining 6000 MNIST datasets. The employed figure of merit to evaluate the suitability of linear regression for this application is the R-score. R-score represents the proportion of the variation in the dependent variable collectively explained by the independent variables. The R-squared score ranges from 0 to 1, with 1 indicating a perfect fit for

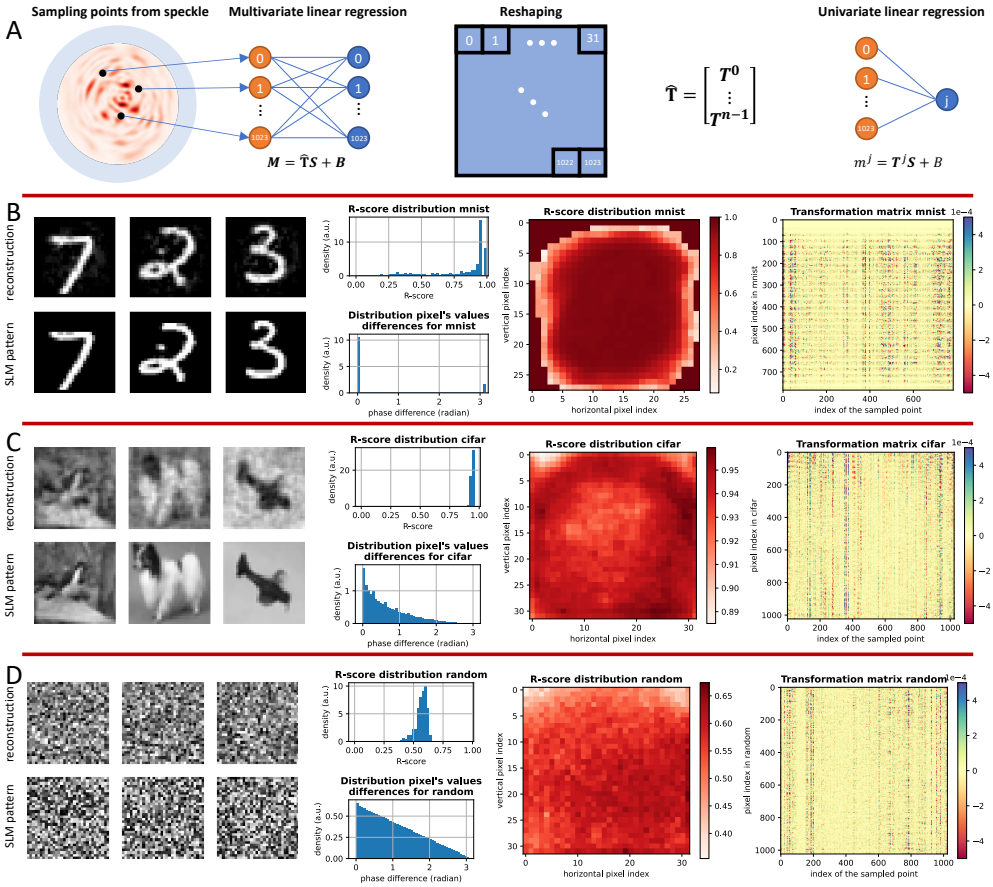


Fig. 2. Panel A: Schematic representation of multivariate linear regression applied to randomly selected points from the speckle image and the pixel values of the Spatial Light Modulator (SLM). The reconstructed SLM pixel values are then reshaped into a two-dimensional array to create a comprehensible image. Alternatively, we can consider each pixel's reconstruction as a univariate problem. Panel B, C, and D: Examples of SLM pattern reconstruction using multivariate regression, the distribution of pixel value differences, the R-score distribution of the linear reconstruction, and the resulting transformation matrix for the MNIST, CIFAR, and Random datasets, respectively. Please note that both the CIFAR and Random image datasets contain images with a size of 32x32 pixels, while the MNIST dataset contains images with a size of 28x28 pixels. This is reflected in their transformation matrices. The transformation matrices for the CIFAR and random images are 1024 by 1024, while the MNIST matrix is 784 by 784.

the linear regression model and 0 indicating a poor fit. The obtained R-score is displayed in its statistical distribution as well as in a 2D array, illustrating the relationship between the quality of reconstruction and the pixel's position within the input image.

Through an analysis of the reconstructed image quality, it becomes evident that linear regression performed relatively well for the MNIST and CIFAR datasets. However, the reconstruction of randomly generated images showed limitations. This trend is also corroborated by the R-score distribution in these three datasets. While both MNIST and CIFAR exhibit an average R-score greater than 0.9, the randomly generated images yield an average R-score of around 0.6. This discrepancy is also visible in the distribution of pixel differences in these datasets. The randomly

generated images exhibit the broadest phase difference distribution. Interestingly, the linear regression technique displays even better performance for the CIFAR dataset compared to MNIST. This is likely due to the fact that pixel value changes in natural scene images are expected to be smoother than those in handwritten digits.

The 2D R-score distribution of these datasets holds critical information about both the dataset itself and the optical setup. In the R-score image of the MNIST dataset, we can identify an area with an R-score of 1. This arises because the digits in MNIST are absent from the corners of the images, meaning the linear regression process requires no effort for reconstruction in those regions. Conversely, an inner boundary area is visible with a weaker R-score (around 0.2), corresponding to the boundary between the white digits and the black background. In this region, where phase differences are pronounced, linear regression struggles because of the condition described by equation (3), leading to statistically weaker reconstruction.

For the CIFAR R-score image, a distinct observation is the presence of relatively weak reconstructions in the two top corners. This arises from the fact that as we approach the image corners, the angle of insertion for each pixel on the SLM gets closer to the Numerical Aperture (NA) cone of the fiber, resulting in diminished reconstruction quality. This behavior has been observed experimentally by physically moving the SLM perpendicular to the optical axis of the system. Additionally, the R-score image of the randomly generated data also displays weaker reconstruction scores in its top two corners. The reason this phenomenon is not apparent in the MNIST dataset reconstruction is that those areas do not contain meaningful information, leading to artificially high R-scores.

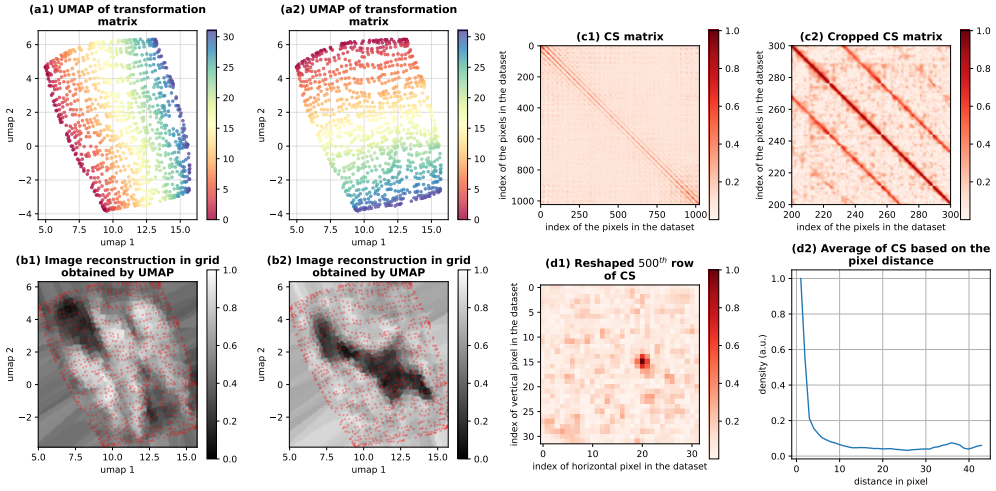


Fig. 3. (a1)-(a2) UMAP projections of the Transformation Matrices in Figure 2 for the CIFAR datasets. The projected points are color-coded based on the horizontal (a1) and vertical (a2) positions of the pixels on the SLM screen. (b1) and (b2) Representation of the reconstructed image using linear regression based on the grid obtained in (a1) and (a2). (c1) and (c2) Cosine similarity matrices comparing the transformation matrix of the CIFAR dataset with a closer look at its details, respectively. (d1) Reshaped version of the 500th row of the cosine similarity matrix in (c1). (d2) The average cosine similarity based on the distances between pixels.

### 2.2.1. The role of the transformation matrix

The transformation matrix **T** links the input and the output of the linear regression networks and it contains information about the learning process. **T** is visualized as as a heat map in Figure 2

(right column), with the horizontal axis indicating the index of the randomly selected point from the speckle, and the vertical axis representing the pixel index on the SLM. By analyzing **T**, in this section we assess how much a linear regression based on the non-locality hypothesis can learn about the the optical system when trained with the different datasets employed in this work.

We employed UMAP to visualize the rows of the transformation matrix, utilizing cosine distance as the metric. The results for the CIFAR transformation matrix are depicted in Figure 3 (a1) and (a2), where each projected row's point is color-coded based on its horizontal and vertical position on the SLM screen, respectively. Interestingly we observed a two-dimensional manifold harmoniously aligned with the grid topology of the SLM. This realization let us hypothesize that the linear regression can glean a substantial amount of information from the underlying physics of the system.

This is supported by multiple evidences. Firstly, the manifold can be used for data visualization purposes. This is exemplified in Figure 3 (b1) and (b2), where the distinct features of a dog and an airplane in the CIFAR dataset reconstruction from Figure 2 are readily discernible through the UMAP-defined points. A second supportive observation is given by how each row of the obtained transformation matrix correlates with one another. This is visualized by the cosine similarity matrix (**CS**) in Figure 3 (c1) and (c2), depicted at two zoom levels. Interestingly **CS** shows several features, apart for the unit diagonal: (i) the main diagonal appears to be thicker than one pixel and (ii) multiple parallel lines to the diagonal. These features refer to the adjacent pixels on the SLM screen. To obtain a clearer evidence, a single row of **CS** can be reshaped to match the size of an image on the SLM screen. This rearrangement portrays the level of similarity between all rows of the transformation matrix and the chosen row, with represented result displayed in Figure 3D-left. Considering that each row of the transformation matrix corresponds to a pixel on the SLM, we can readily observe that the pixels in close proximity to the selected pixel (which is the one with maximum value) exhibit notably high similarity. Additionally, we can compute the average value of cosine similarity based on the physical distances between pixels. This result is depicted in Figure 3D-right. The graph illustrates that when the distance between pixels exceeds four pixels, their cosine similarity index decreases significantly, approaching nearly zero. This trend signifies effective perpendicularity between those distant pixels.

This highlights how the transformation matrix has learned an effective range for interactive pixels on the SLM screen despite that it is constructed on a subset of speckle patterns. Together with the obtained 2D manifold, this supports the hypothesis that the linear regression model based on non-locality can gain insights from specific properties of the optical system.

The efficiency at which this happens depends on the extent at which the condition in equation (3) is verified. For instance for the MNIST dataset, which features strong phase variations between the digit and the background, the cosine similarity matrix show less prominent similarity features, e.g. a thinner diagonal as well as a lower number of diagonal lines (Supplementary Figure 5). The random dataset, on the other hand, which has highest possible randomness of phase variations, show no similarity features in **CS**. For both these datasets the 2D manifold of the SLM screen through the transmission matrix cannot be retrieved (Supplementary Figure 6). All this results in the lower R-score values found for MNIST and random datasets in Figure 2, highlighting how the processing of datasets not fullfilling the linearity criteria in equation (3) should be demanded to advanced deep learning models. In the following we introduce a non-locality based global attention mechanism that allows achieving high performance and high fidelity reconstruction also in the case of random dataset.

### *2.3. Enhanced image transmission fidelity with global attention*

Although classical CNN can handle datasets featuring the above-described non-linearity, they are deceived by non-locality. In this section we first describe this limitation highlighting how the presence of non-locality makes CNN operating on features extraction rather than on pixel-by-pixel

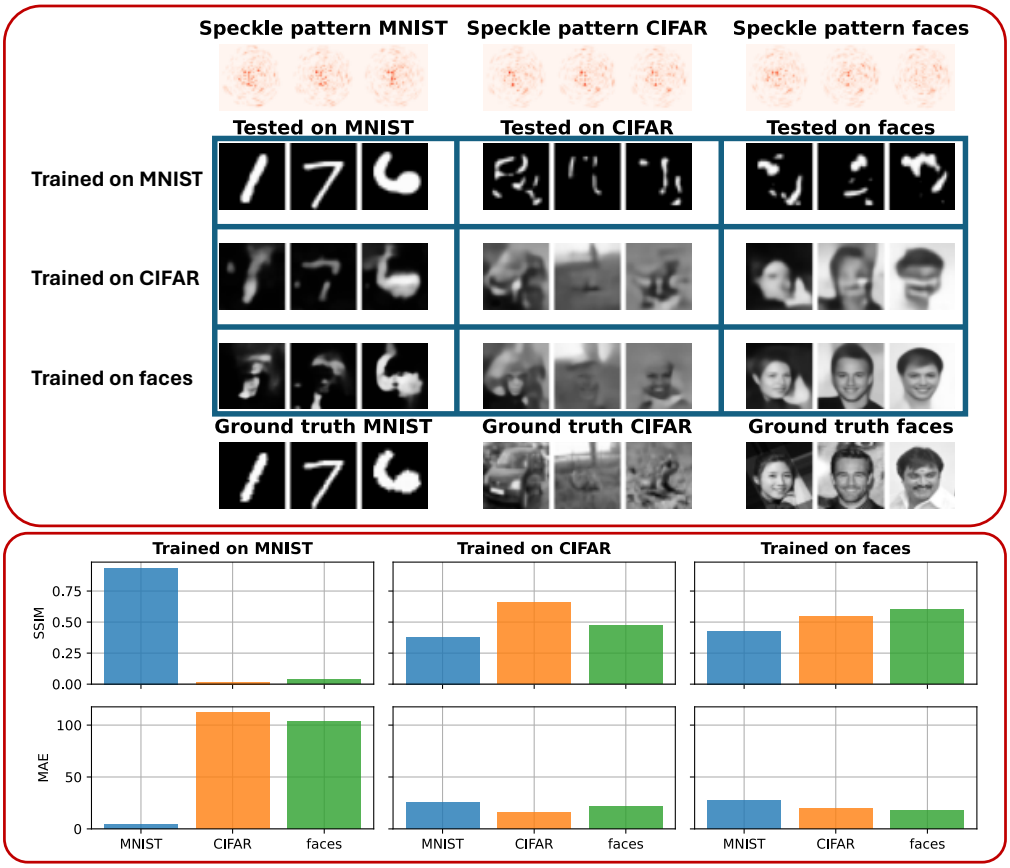


Fig. 4. Comparing the reconstruction performance of a network trained on three distinct datasets—specifically MNIST, CIFAR, and celebrity faces. A subset of each dataset is separated before training for testing purposes. The celebrity faces dataset used for this analysis was obtained from the source [29].

reconstruction of the transmitted grid topology data. We then introduce a novel global attention mechanism that makes CNN operating efficiently in the presence of both non-locality and non-linearity in the problem.

Classically employed CNN architectures are based on U-Net, owing to their established capability for image-to-image transformations in tasks like medical image segmentation. However, CNN-based architectures are constrained by their filter sizes when capturing non-local patterns within grid topology data. While using larger filter sizes can enhance performances , it also comes at the cost of increased computational demands. Another avenue to augment their effectiveness involves increasing the depth of the network, enabling deeper layers to gain a more comprehensive understanding of global patterns in the image data. To describe CNN limits we re-implemented a U-Net architecture incorporating nested residual networks in each stage of both the encoder and decoder sections (ResUNet, full network in Supplementary Figure 7). The inclusion of residual networks assists in achieving greater network depth while mitigating issues like accuracy saturation and gradient vanishing. For training, we employed the mean absolute error (MAE) as the cost function. This choice ensures that the training criterion centers around pixel-wise reconstruction, in contrast to other loss functions like the structural similarity index (SSIM), which consider image structure and features. The network is fed by the speckle patterns and

the outcomes of image reconstruction are displayed in Figure 4. The network was trained three times using distinctive datasets—MNIST, CIFAR-10, and celebrity faces [29]. From each dataset, 50,000 images were considered, with the first 40,000 for training, the next 5,000 for validation, and the final 5,000 for testing. The celebrity faces dataset consists of images of celebrities with pixel dimensions of 128x128. This increased freedom in image size compared to the number of guiding modes in the fiber (approximately 1600) suggests a likelihood of information loss during the reconstruction process.

Figure 4 illustrates the performance of each trained network when tested with different datasets. The network trained on MNIST excels at reconstructing MNIST images but struggles when applied to the testing set of other datasets, revealing a lack of generality. This behavior arises because deep learning methods for image reconstruction tend to rely on image features rather than pixel-wise reconstruction. The CIFAR-trained network performs significantly better than the MNIST counterpart in terms of generalization, likely due to the absence of obvious image features in CIFAR compared to MNIST. As for the celebrity faces dataset, resembling MNIST in terms of distinctive features, the network's performance is hindered by its exclusive focus on facial features. An intriguing observation with the celebrity faces network is that while the reconstructed faces are visually appealing, they represent different individuals with respect to the ground truth. While some features like facial orientation are reconstructed, others such as ethnicity, lips' shape, or background are not. This becomes more evident when the network is tested on MNIST and CIFAR datasets, where facial features emerge in both digits and CIFAR image reconstructions.

Therefore addressing this feature-based reconstruction challenge could involve using feature-less datasets in the training process, such as randomly generated images used in the linear regression problem. However our experiments reveal that the ResUNet network struggles to perform well with feature-free datasets (see red line in the benchmark in Figure 5A). Indeed, in ResUNet the input image undergoes a series of convolution operations, each generating a feature map from the input. These feature maps are then stacked in the third dimension (channel dimension) to create a three-dimensional tensor (excluding the batch dimension). This makes the features extracted constrained by the used filter size.

To enable the ResUNet network to take advantage of non-locality in general datasets we introduced a global awareness module (GAM). The GAM is added to the longest skips connection in the ResUNet algorithm (Figure 5A), enabling significantly high performances with respect to other employed models, while using a much smaller number of parameters (see benchmark in Figure 5A).

Our implementation of GAM is depicted in Figure 5B: it develops in upper and lower pathways, both exploiting non-locality. In the upper pathway, a single two-dimensional convolution operator with a filter size of (1x1) is applied to the tensor, resulting in a 2D feature map. It's important to note that this technique involves weighted averaging across the channels in the tensor. As observed from the linear regression results, not all components of this 2D feature map are fully independent, especially when the feature map size significantly exceeds the intended SLM pattern. For instance, in the CIFAR dataset, where the SLM screen data size is 32x32 pixels and the recorded speckle's size is 128x128, a scenario akin to the linear regression case arises, enabling random selection of specific points while ignoring others. To achieve this, a custom layer is designed to identify indices of the selected components in the feature map, converting them into a one-dimensional array. This reduced array then passes through a single-layer MLP with linear activation, yielding the first linear embedding (LE1 in Figure 5) derived from our input tensor. Notably, as the network delves deeper, this feature size converges with the intended SLM pattern, rendering input reduction unnecessary. In fact, this reduction aims to streamline the complexity of the linear embedding technique.

In the lower pathway, the three-dimensional tensor is permuted (rotated) to allow another

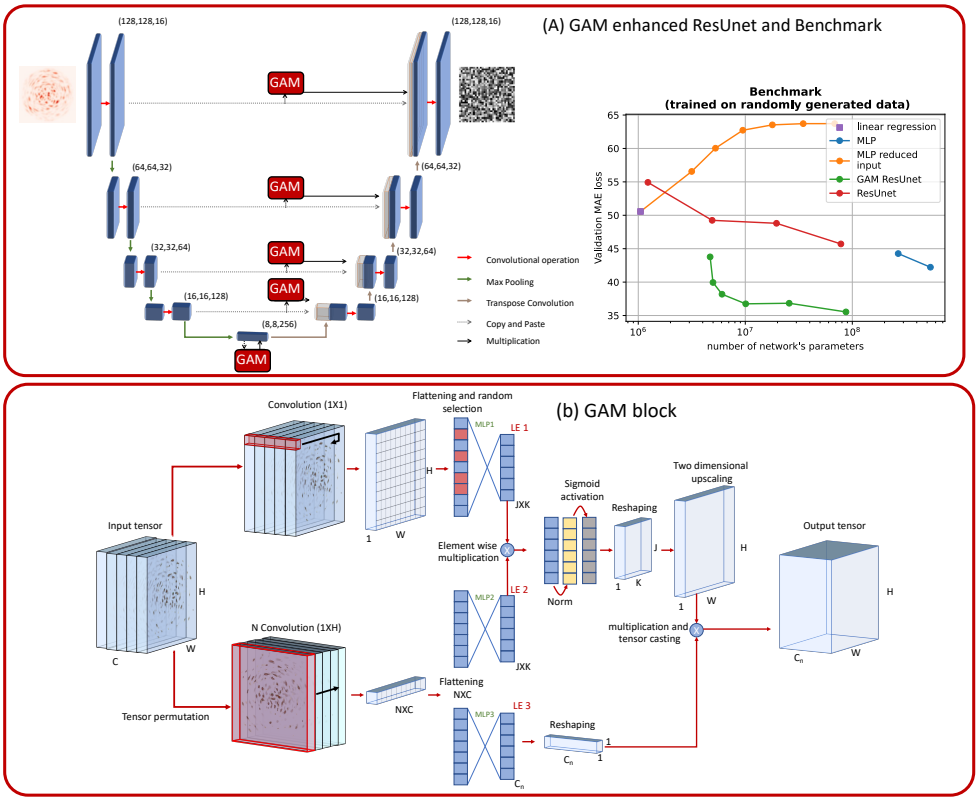


Fig. 5. Panel A illustrates how the GAM block can be incorporated into the ResUNet architecture. It also demonstrates how the proposed network operates in comparison to other architectures, particularly when networks with varying complexities are considered. Panel B: Schematic of the GAM block. The upper path extracts a linear embedding of the input tensor using a 1x1 convolution filter. Subsequently, a portion of the data is randomly selected and passed through a single-layer fully connected layer (MLP1) with linear activation, resulting in the first linear embedding (LE1). On the bottom path, the input tensor is initially permuted to enable convolutional filters to traverse across channels. Following this, (N) convolutional filters with a size of 1xH are applied. The output then passes through MLP2 and MLP3, resulting in two additional linear embeddings of the input tensor, namely LE2 and LE3. LE1 and LE2 undergo element-wise multiplication and, after normalization, pass through a sigmoid activation function, constraining the output of each node between 0 and 1. Subsequently, after reshaping into a two-dimensional tensor and scaling, it is multiplied and casted to the reshaped version of LE3, yielding the output tensor.

convolution operation to traverse the channels. Unlike the previous pass, where the 1x1 convolution layer computed channel averages, in this phase, the convolution aims to directly extract non-local information from the channels. The size of this convolution operator is carefully chosen to encompass information from all elements within the feature map. The number of these filters remains a hyperparameter within this technique. This operation directly yields a one-dimensional tensor as its output. This output subsequently passes through two MLP networks, resulting in two linear embeddings of the input tensor—designated as LE2 and LE3 in Figure 5B. LE1 and LE2 undergo element-wise multiplication, followed by normalization and

softmax activation. The results are then reshaped into a two-dimensional tensor. Size equality between LE1 and LE2 is crucial for valid element-wise multiplication, and their dimensions are considered hyperparameters of this technique. The reshaped two-dimensional tensor is upscaled to match the feature map size of the original three-dimensional input tensor. However, LE3 is reshaped to form a three-dimensional tensor with a 1x1 feature map size, and this tensor is then multiplied with the upscaled tensor. This multiplication and casting result in a three-dimensional tensor identical in feature map size to the input tensor, but with the channel size determined by LE3 — another hyperparameter of this technique.

The input tensor's size for the GAM operator varies based on the skip connection it receives as input. For the top skip connection in the ResUNet architecture, the feature map size is 128x128. As a result, a sampling operation from the two-dimensional feature map is employed to mitigate complexity and reduce the parameters in MLP1. LE1 can also match the SLM pattern's size in this context—32x32. This choice, in turn, determines LE2's size, given that both LE1 and LE2 share identical dimensions. However, as the ResUNet deepens, feature map sizes decrease due to max-pooling layers in the encoder section, rendering random data point selection in the reduced tensor increasingly unnecessary. Of note, the output of GAM undergoes element-wise multiplication with the concatenation of the skip connection and the previous decoder layer's upscaling in ResUNet.

To demonstrate the effectiveness of the aforementioned technique in enhancing ResUNet performance, we conducted a benchmark using randomly generated images as training daset. For each network in this benchmark, we implemented 20 training epochs, maintaining consistent network hyperparameters including batch size and learning rate. The results are depicted in Figure 5 (top-right), where the x-axis logarithmically represents the number of parameters in each network. The y-axis illustrates the minimum validation set loss, presented as the mean absolute error.

The ResUNet used for data in Figure 4 is represented by the red line. This network's parameter count increased by doubling the number of filters in each layer, commencing with 8 filters in the initial ResUNet layer. The ResUNet enhanced by the GAM block is shown in green. Similar to ResUNet, The parameter count enhancement in the GAM-enhanced model was achieved by simply adding more filters to the convolutional operators. Each addition doubled the previous count. However, in this case, it began with 2 filters in the initial layer and then doubled up. Notably, the complexity and parameter count of the GAM block remained constant for the sake of simplicity of the comparison. The results reveal that the GAM-enhanced ResUNet outperforms ResUNet, even with more than ten times smaller parameter size. For the sake of completeness we have represented in the same graph the results obtained for the linear regression method described in previous section (magenta dot). Interestingly, it outperforms the ResUNet for the same number of parameters. However, increasing the number of layers (e.g. transforming the linear regression into a multilayer perceptron with a reduced input) decreases the effectiveness of the method (orange line). GAM-ResUNet also improves the performances with respect to the MLP, similar to the one in ref [15], which resulted to be the closest to our non-locality awareness approach in terms of loss but employing a more than 100 times larger network.

### *2.4. Transfer learning and pixel wise reconstruction*

Now equipped with a network that exhibits relatively strong reconstruction performance on randomly generated datasets, we can leverage its capabilities to transfer pixel-wise reconstructions from these randomly generated images to other datasets. For this we propose a post-training phase, in which the trained network on randomly generated datasets continues its training with additional data on a specific dataset, such as CIFAR and celebrity faces.

The outcomes of this transfer learning are depicted in Figure 6 (with more examples in the supplementary figures 8 and 9). Interestingly, even before the post-training process on any of

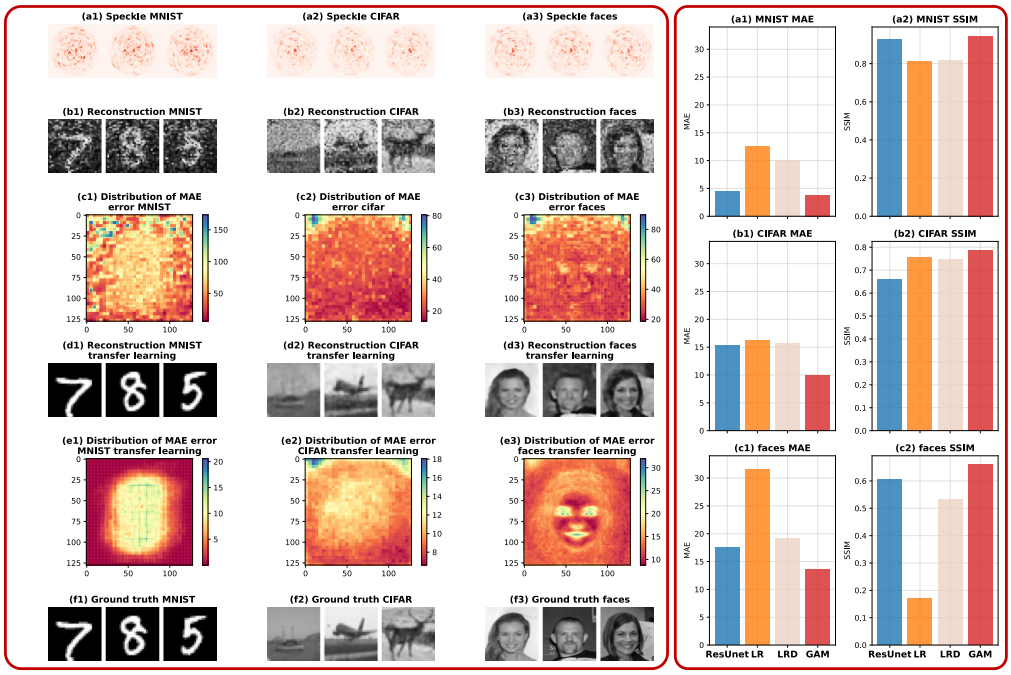


Fig. 6. Panel A displays speckle pattern examples from the testing sets of MNIST, CIFAR, and faces in (a1), (a2), and (a3), respectively. The reconstructions of these testing speckle using the proposed network, trained on a random dataset, are shown in (b1), (b2), and (b3). (c1), (c2), and (c3) present the Mean Absolute Error (MAE) distribution for the entire testing data in MNIST, CIFAR, and faces, respectively. (d1), (d2), and (d3) showcase the reconstructions of the same data in (a1)-(a3) when the network, initially trained on a random dataset, undergoes post-training on MNIST, CIFAR, and faces, respectively. (e1), (e2), and (e3) display the MAE error distribution for the entire testing datasets of MNIST, CIFAR, and faces, respectively. Lastly, (f1), (f2), and (f3) represent the ground truth for the presented reconstructions. Panel B illustrates the quantitative performance of the post-trained network and compares it to Res-UNET and the presented linear regression. It's important to note that to enhance the results of linear regression, we applied a smoothing filter to mitigate noise in the linear regression output, serving as an additional improvement for a fairer comparison. Additionally, higher SSIM values indicate better performance, while lower MAE values signify improved performance. The celebrity faces dataset used for this analysis was obtained from the source [29].

these datasets, the network demonstrated a relative capability to reconstruct the overall image context. A notable example is evident with MNIST, where all the individual digits are distinctly recognizable. Importantly, the distribution of loss across all reconstructed pixels does not reveal any dataset-specific information. This emphasis on pixel-wise reconstruction highlights the network's ability to retain its focus on feature-based reconstruction rather than capturing specific dataset patterns.

Following the second training step, the quality of reconstruction significantly improves. Of particular significance, facial reconstructions are now representing the exact individuals – a striking contrast to the results of the ResUNet model, where the reconstructed faces enhanced fidelity but did not resemble the ground truth (Figure 4). The slight blurriness in the facial reconstructions is a result of the inherent limitation that the fiber carries fewer modes than

number of pixels in faces dataset. This introduces a physical constraint that hampers the complete transmission of information.

To further demonstrate how the second training step enhances feature-based reconstruction over pixel-wise reconstruction, we calculated error distributions across all reconstructed pixels. The results compellingly illustrate that the error distribution correlates with distinctive features unique to each dataset, showcasing the added feature-based reconstruction. For instance, the errors in celebrity face reconstructions predominantly align with discrepancies in facial features like eyes and facial boundaries. This effect is observed while a significant portion of other information, such as background details and overall facial structure, is reconstructed more accurately.

A comprehensive quantitative assessment of the proposed method, in comparison to other techniques introduced in this study, is illustrated in Figure 6. This visualization provides a clear demonstration of how our method outperforms ResUNet across all tested datasets, both in terms of image fidelity and Mean Absolute Error (MAE) loss. Notably, the performance of our linear regression models in the CIFAR dataset is relatively strong, resulting in outcomes comparable to ResUNet. For all cases of reduced input linear regression, an image denoising step was applied to enhance the quality of the reconstructed images, as linear regression inherently lacks denoising capabilities. This denoising step significantly improved the output of linear regression for all datasets.

Furthermore, we computed the Structural Similarity Index (SSIM) for the outputs of all employed methods. Once again, the proposed method exhibited better performance. Another noteworthy observation arises when applying image denoising to the faces reconstructed through linear regression. This additional step leads to a substantial improvement in SSIM index compared to the original linear regression output.

However, the excellence of the proposed method goes beyond numerical comparisons. Notably, the reconstructed faces from ResUNet, despite achieving acceptable quantitative measures such as MAE and SSIM, are not visually convincing and do not faithfully represent the individuals depicted in the SLM screen (Figure 4). In stark contrast, our proposed method effectively overcomes this limitation, resulting in highly accurate facial reconstructions that closely resemble the actual individuals (Figure 6A).

### 2.5. Amplitude and phase Modulation

With a small modification, the entire presented method can be applied to the more general case of amplitude and phase modulation. The equation for the combined modulation reads as follows:

$$\boldsymbol{E}(x,y) = \sum_{n} A(\phi_n)\boldsymbol{U}_n(x,y)e^{j\beta(\phi_n)} \tag{7}$$

where $A(\phi_n)$ represents amplitude modulation, and $\beta(\phi_n)$ represents phase modulation. If we consider small deviations for both of these functions around a specific modulation point $\phi_0$ ($\phi_n = \phi_0 + \epsilon_n$), we can approximate both amplitude and phase modulation functions as $A(\phi_n) \approx A(\phi_0) + A^{'}(\phi_0)\epsilon_n$ and $\beta(\phi_n) \approx \beta(\phi_0) + \beta^{'}(\phi_0)\epsilon_n$, respectively. Here, $A^{'}(\phi_0)$ and $\beta^{'}(\phi_0)$ denote the first derivatives of $A$ and $\beta$ with respect to $\phi$ at the $\phi_0$ point.

By using the Taylor expansion of the exponential function in equation 7 and rewriting the formula for the intensity of the electric field, we obtain the following expression:

$$\begin{aligned} |\boldsymbol{E}(x,y)|^2 = \sum_{n}\sum_{m}(A^2(\phi_0) + \epsilon_n(A^{'}(\phi_0)A(\phi_0) + jA^2(\phi_0)\beta^{'}(\phi_0)) \\ +\epsilon_m(A^{'}(\phi_0)A(\phi_0) - jA^2(\phi_0)\beta^{'}(\phi_0))\boldsymbol{U}_n(x,y).\boldsymbol{U}_m^*(x,y) \end{aligned} \tag{8}$$

This relationship indicates a clear linear correlation between the modulating factor ($\epsilon_n$) and the intensity of the electric field. Similar to the phase-only case, this correlation holds as long as

the value of $(\epsilon_n)$ is relatively small.

In this context, we applied both linear regression and GAM methods to analyze the data provided in the study by Caramazza et al. (2019) on transmission characteristics. The experimental setup utilized a one-meter-long step index fiber with a core size of 105 $\mu m$, capable of carrying approximately 9000 optical modes. The researchers imprinted images onto the fiber by representing the intensity of the electric field using an SLM, a polarizing beam splitter (PBS), and a half-wave plate.

The dataset employed in their experiment consisted of 45,000 images with a resolution of $92 \times 92$ pixels sourced from the ImageNet dataset, featuring natural scenes like plants and animals. This specific image selection aimed to avoid apparent geometric features found in handwritten digits of the MNIST dataset. The deliberate choice of this dataset positions it as an ideal candidate for our GAM method, as we aspire to achieve pixel-wise reconstruction without the two-step training process outlined in the previous section. This approach eliminates the need for transfer learning and allows us to leverage the network's pixel-wise reconstruction capabilities in a single-step training process. However, the testing set is considered to be part of the Muybridge motion data, comprising sequences featuring a horse, parrot, punch, and a cat.

However, due to the optical setup in Caramazza et al.'s method involving more modes than our setup, we introduced slight modifications to both linear regression methods and GAM. Specifically, for the case of the linear regression, we randomly selected 8468 points (equivalent to $92 \times 92$ pixels) from the speckle. For the GAM method, we expanded the dimensionality of the randomly selected points in MLP1 of Figure 5 to 2500, where random selection is necessary. It's important to note that as the size of the feature map decreases in the deeper layers of the network, random selection becomes unnecessary.

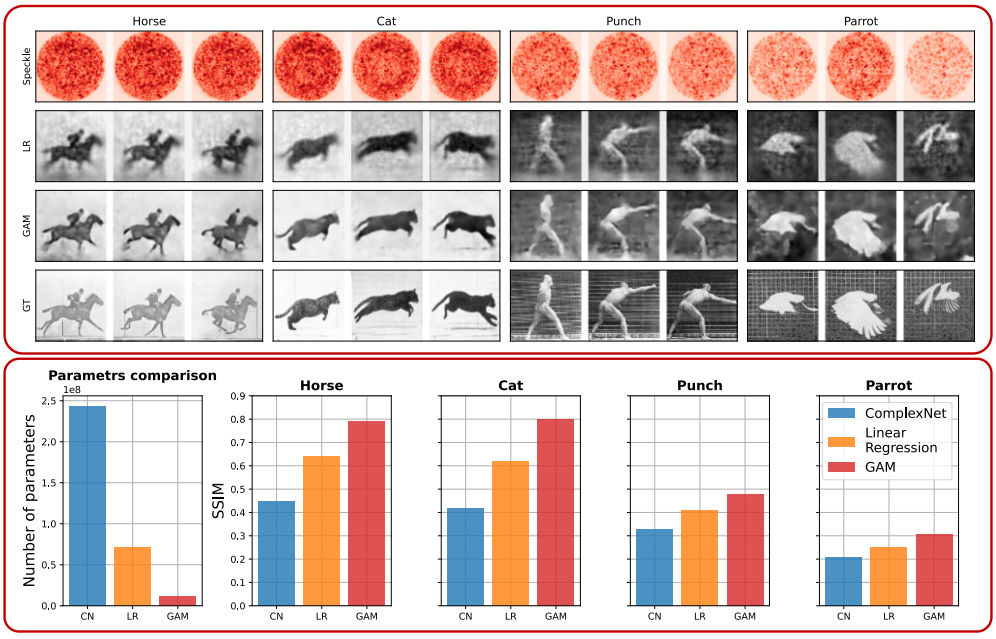


Fig. 7. (Top) Reconstruction examples featuring linear regression (LR), GAM, and their corresponding ground truth. (Bottom) Parameter comparison among ComplexNet [16], linear regression, and GAM. Additionally, the SSIM comparison evaluates the reconstructed images against the ground truth for each testing set separately when utilizing ComplexNet, linear regression, and GAM for image reconstruction. The data for this analysis was obtained from the source [16].

Figure 7 (top) showcases the reconstructed samples obtained through both linear regression and GAM techniques. In-depth quantitative comparisons among linear regression, GAM, and the ComplexNet proposed in [16] are presented in Figure 7 (bottom). Remarkably, both GAM and linear regression exhibit superior reconstruction performance while utilizing a significantly lower number of parameters compared to the ComplexNet models proposed in the referenced work.

It is noteworthy that the GAM, in particular, stands out by employing 25 times fewer network parameters than the ComplexNet model. Despite this substantial reduction in parameter count, the GAM achieves the best performance, surpassing both linear regression and ComplexNet models. Specifically, the GAM outperforms the ComplexNet by approximately two times in terms of Structural Similarity Index (SSIM), highlighting its remarkable efficiency and effectiveness in image reconstruction.

## 3. Discussions and conclusion

In the preceding sections, we demonstrated that leveraging a non-locality notion arising from the light propagation in MMFs enables the development of a novel global attention mechanism. This mechanism not only significantly enhances the performance of deep convolutional neural networks in image reconstruction through the produced speckles, but it also facilitates pixel-wise reconstruction, contributing to the accurate reconstruction of the transmitted data.

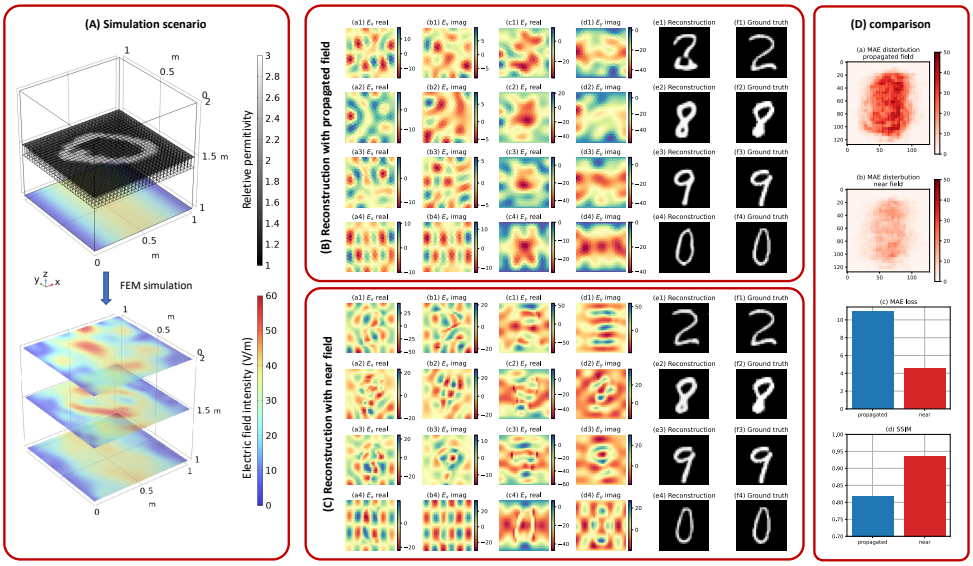


Fig. 8. pannel A show the schematic of the simulated scenario using FEM algorithm. top, shows the a rectangular waveguide boundaries with perfect electrical conductor at its boundaries. The fundamental mode is propagating from the bottom of the structure and hits a material distribution with varying relative permitivity between 1 to 3 as shown. bottom shows the norm of electric field distribution on the excitation port, near the material distribution and one meter away from it. panel B and C show the transverse electric field components and the material distribution reconstruction based on them for both propagated(far) and near fields, respectively. panel D show the reconstruction fidelity of obtained from the near and far field.

The GAM-enhanced method can find a wide applicability in scattering problems, since in photonic systems scattering often induce a non-locality in the input-output relationship, which can compensate data loss. Such a problem is known as inverse design as one is trying to find a

geometry of the problem based on the given electromagnetic response. To give an example on this respect, we have conducted data-driven analysis on the interaction of electromagnetic waves with material distributions, followed by their subsequent scattering and coupling in a guiding system which can account for a much smaller number of modes with respect to MMFs. In this instance, we utilized finite element (FEM) simulations to delve into the subject in the simulation scenario depicted in Figure 8A. At the upper part of this panel, a permittivity distribution is positioned within a perfect electric conductor waveguide. The waveguide is excited using its fundamental mode from the structure's bottom. For this square cross-sectional waveguide, with a side length of 1 meter, it can be determined that 40 distinct modes are capable of propagation. Once the fundamental mode encounters the permittivity distribution, the resulting scattered wave then triggers the excitation of other waveguide modes in both directions. This occurrence leads to the generation of an electric field distribution, as shown at the bottom of Figure 8B. The electric field distribution at the input port, near the permittivity distribution, and half a meter away from it is illustrated. Evidently, the back reflected wave distorts the electric field distribution at the input port. Additionally, both the near field and the propagated field experience distortion, carrying information from the material distribution.

Importantly, this problem is inherently different from our previous investigation on MMFs, where the propagating medium remained constant while the source varied. In contrast, in this case, the propagating medium can change by adjusting the permittivity distribution, while the excitation source remains fixed. Still non-local effect arise due to potential coupling between distinct segments of the material distribution and, akin to the MMFs problem, there is an added non-locality due to the interaction between the different modes after the scattering events.

Figures 8, Panel B and C, exhibit the reconstruction of the permittivity distribution based on the electric field distribution in the propagated field (0.5 meters from the material distribution) and the near field. In this case, both transverse electric field components are utilized, as knowledge of their profiles guarantees comprehension of the electric field across other cross sections of the waveguide, following the principles of electrodynamic theory. Given that four different field distributions are now under consideration —comprising the real and imaginary parts of both the x and y electric field distributions— adaptations are required in the network architecture with respect to the MMFs case. These two-dimensional field distributions are vertically stacked into the third dimension, resulting in a unified three-dimensional input. Our GAM-enhanced method adeptly processes this three-dimensional input and endeavors to map it to the permittivity distribution throughout the learning process. It is evident that the reconstruction of the near field surpasses that of the propagated field. This disparity stems from the fact that the waveguide can only accommodate 40 propagating modes, resulting in information loss as the wave propagates and leading to a less precise reconstruction. Similar to prior challenges, this reconstruction is anchored in MNIST's features, acknowledging the limited generality of the trained network. Yet, there exists potential to improve generalization by running simulations using randomly generated material distributions.

The central objective of this scattering endeavor revolves around identifying an inverse transformation for the FEM simulation. Such an inverse design bears relevance in numerous realms of electromagnetic engineering, including the design of metamaterials and solving inverse scattering problems. These scenarios also involve non-local effects [30], and for instance can be extended to the case of light propagation in highly scattering media. This is the case of the brain, with neural endoscopy applications being limited by a very short photons mean free path ($<20\mu m$ in some regions). The here-presented GAM-enhanced method can therefore represent a complementary tool for engineering brain imaging methods, and can find applications in the more general context of biological tissues where scattering hinder light propagation more than direct absorption of photons.

## 4. Data and Code Availability

For the phase-only modulation, this study utilized 120,000 randomly generated images along with their corresponding speckle patterns. Additionally, for each of the three datasets (faces, CIFAR, and MNIST), we obtained pairs of 10,000 data points.

The dataset for phase-only modulation and the code example for reproducing the results of the GAM and linear regression on ImageNet data in [16], are publicly available at the time of publication.

## 5. Funding

M.K. and L.C. contributed equally and are co-first authors of this work. M.D.V. and Fe.P. jointly supervised and are co-last authors of this work. L.C., M.D.V., and Fe.P. acknowledge funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 828972. L.C., M.D.V. and Fe.P. acknowledge funding from the Project "RAISE (Robotics and AI for Socio-economic Empowerment)" code ECS00000035 funded by European Union – NextGenerationEU PNRR MUR - M4C2 – Investimento 1.5 - Avviso "Ecosistemi dell'Innovazione" CUP J33C22001220001. M.D.V., and Fe.P. acknowledge funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 101016787. M.D.V. acknowledges funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 692943. M.D.V., and Fe.P. acknowledge funding from the U.S. National Institutes of Health (Grant No. 1UF1NS108177-01). M.D.V. and Fe.P. acknowledge funding from European Research Council under the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 966674. Fe.P. acknowledges funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 677683.

## 6. Disclosures

M.D.V. and Fe.P. are founders and hold private equity in Optogenix, a company that develops, produces and sells technologies to deliver light into the brain. Tapered fibers commercially available from Optogenix were used as tools in the research. MDV: Optogenix srl (I). FP: Optogenix srl (I).

## References

1. I. M. Vellekoop and A. Mosk, "Focusing coherent light through opaque strongly scattering media," Opt. letters **32**, 2309–2311 (2007).
2. R. Di Leonardo and S. Bianchi, "Hologram transmission through multi-mode optical fibers," Opt. Express **19**, 247–254 (2011).
3. I. N. Papadopoulos, S. Farahi, C. Moser, and D. Psaltis, "Focusing and scanning light through a multimode optical fiber using digital phase conjugation," Opt. Express **20**, 10583–10590 (2012).
4. M. C. Velsink, Z. P. Lyu, P. W. H. Pinkse, and L. V. Amitonova, "Comparison of round- and square-core fibers for sensing, imaging, and spectroscopy," OPTICS EXPRESS **29**, 6523–6531 (2021).
5. S. H. Li, C. Saunders, D. J. Lum, *et al.*, "Compressively sampling the optical transmission matrix of a multimode fibre," LIGHT-SCIENCE & APPLICATIONS **10** (2021).
6. T. Čižmár and K. Dholakia, "Shaping the light transmission through a multimode optical fibre: complex transformation analysis and applications in biophotonics," Opt. Express **19**, 18871–18884 (2011).
7. L. Collard, F. Pisano, D. Zheng, *et al.*, "Holographic Manipulation of Nanostructured Fiber Optics Enables Spatially-Resolved, Reconfigurable Optical Control of Plasmonic Local Field Enhancement and SERS," Small **n/a**, 2200975 (2022).
8. 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," Opt. Lett. **43**, 419–422 (2018).
9. D. Stellinga, D. B. Phillips, S. P. Mekhail, *et al.*, "Time-of-flight 3d imaging through multimode optical fibers," Science **374**, 1395–1399 (2021).

10. L. Collard, F. Pisano, M. Pisanello, *et al.*, "Wavefront engineering for controlled structuring of far-field intensity and phase patterns from multimodal optical fibers," APL Photonics **6**, 51301 (2021).
11. I. T. Leite, S. Turtaev, X. Jiang, *et al.*, "Three-dimensional holographic optical manipulation through a high-numerical-aperture soft-glass multimode fibre," Nat. Photonics **12**, 33–39 (2018).
12. S. A. Vasquez-Lopez, R. Turcotte, V. Koren, *et al.*, "Subcellular spatial resolution achieved for deep-brain imaging in vivo using a minimally invasive multimode fiber," Light. science & applications **7**, 110 (2018).
13. S. Turtaev, I. T. Leite, T. Altwegg-Boussac, *et al.*, "High-fidelity multimode fibre-based endoscopy for deep brain in vivo imaging," Light. Sci. & Appl. **7**, 92 (2018).
14. N. Borhani, E. Kakkava, C. Moser, and D. Psaltis, "Learning to see through multimode fibers," Optica **5**, 960–966 (2018).
15. B. Rahmani, D. Loterie, E. Kakkava, *et al.*, "Actor neural networks for the robust control of partially measured nonlinear systems showcased for image propagation through diffuse media," Nat. Mach. Intell. **2**, 403–410 (2020).
16. P. Caramazza, O. Moran, R. Murray-Smith, and D. Faccio, "Transmission of natural scene images through a multimode fibre," Nat. communications **10**, 2029 (2019).
17. C. Zhu, E. A. Chan, Y. Wang, *et al.*, "Image reconstruction through a multimode fiber with a simple neural network architecture," Sci. reports **11**, 896 (2021).
18. R. Xu, L. Zhang, Z. Chen, *et al.*, "High accuracy transmission and recognition of complex images through multimode fibers using deep learning," Laser & Photonics Rev. **17**, 2200339 (2023).
19. B. Rahmani, D. Loterie, G. Konstantinou, *et al.*, "Multimode optical fiber transmission with a deep learning network," Light. science & applications **7**, 69 (2018).
20. L. Collard, M. Kazemzadeh, L. Piscopo, *et al.*, "Exploiting holographically encoded variance to transmit labelled images through a multimode optical fiber," arXiv preprint arXiv:2309.15532 (2023).
21. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18,* (Springer, 2015), pp. 234–241.
22. S. Resisi, S. M. Popoff, and Y. Bromberg, "Image transmission through a dynamically perturbed multimode fiber by deep learning," Laser & Photonics Rev. **15**, 2000553 (2021).
23. P. Fan, T. Zhao, and L. Su, "Deep learning the high variability and randomness inside multimode fibers," Opt. express **27**, 20241–20258 (2019).
24. G. Wu, Y. Sun, L. Yin, *et al.*, "High-definition image transmission through dynamically perturbed multimode fiber by a self-attention based neural network," Opt. Lett. **48**, 2764–2767 (2023).
25. A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," Adv. neural information processing systems **30** (2017).
26. Y. Chen, B. Song, J. Wu, *et al.*, "Deep learning for efficiently imaging through the localized speckle field of a multimode fiber," Appl. Opt. **62**, 266–274 (2023).
27. Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision,* (2021), pp. 10012–10022.
28. B. Song, C. Jin, J. Wu, *et al.*, "Deep learning image transmission through a multimode fiber based on a small training dataset," Opt. express **30**, 5657–5672 (2022).
29. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision,* (2015), pp. 3730–3738.
30. J. R. Capers, S. J. Boyes, A. P. Hibbins, and S. A. Horsley, "Designing the collective non-local responses of metasurfaces," Commun. Phys. **4**, 209 (2021).