nature photonics

Supplementary information

https://doi.org/10.1038/s41566-025-01657-6

Model-free estimation of the Cramér–Rao bound for deep learning microscopy in complex media

In the format provided by the authors and unedited

CONTENTS

S1. OPTICAL SETUP

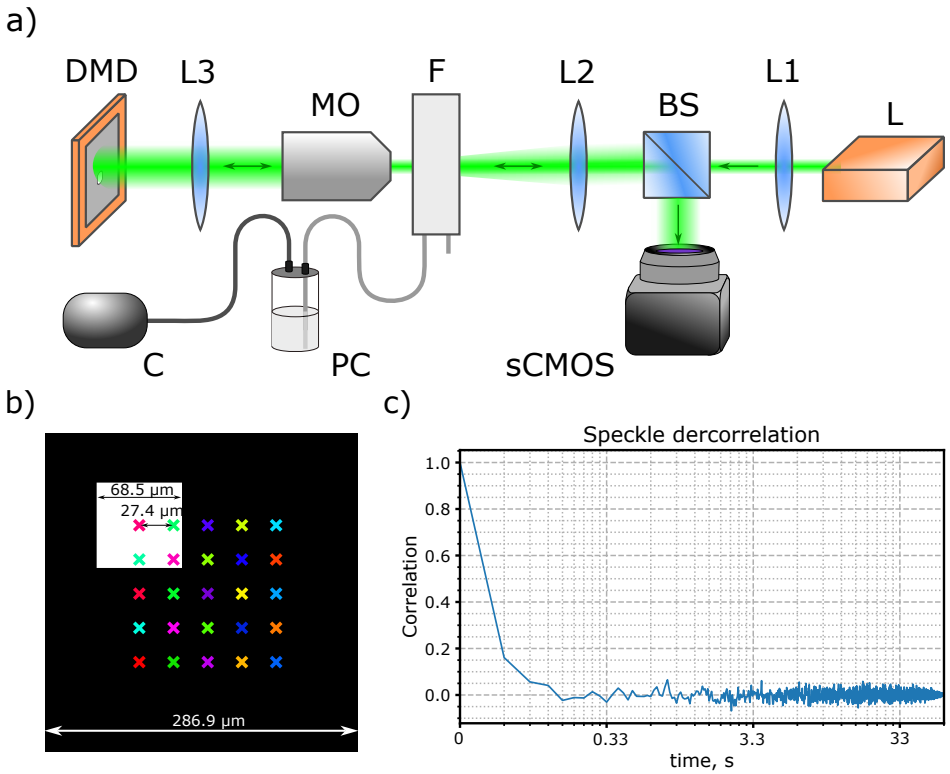

FIG. S1. (a) Experimental setup: L - laser diode (Thorlabs DJ532-40), L1, L2, L3 - lenses (focal length $f = 100\,\text{mm}$), BS - 50/50 beam splitter, F - flow cuvette (Helma, 6.2 µL, optical path length 100 µm), DMD - digital micromirror device (Vialux V-7001), MO - microscope objective (Olympus Plan N 10x), PC - pressure chamber, C - compressor, sCMOS - camera (Andor Zyla 5.5). (b) Sketch representing the dimensions of the target (in white), of the field of view (in black) and of the displacements used for the test set (colored crosses). The target is composed of 5×5 DMD pixels and is displaced by 2 pixels between two adjacent positions. Dimensions are given here in the DMD plane (a magnification factor of ×0.18 should be applied to calculate corresponding dimensions in the sample plane). (c) Normalized intensity correlation function of the speckles measured in the absence of the target. The observed decorrelation time is around 30 ms, which is of the order of the time interval between successive frames (33.3 ms) and much longer than the exposure time of the camera (200 µs).

S2. ESTIMATING THE TARGET POSITION FROM BALLISTIC AND SCATTERED LIGHT

For small optical thicknesses, we essentially measure the ballistic light coming from the target: the target position is visible on single-shot intensity images [Fig. S2(a)], and the average intensity collected at the target position is significantly larger than the random fluctuations caused by the scattered light [Fig. S2(c) and Fig. S2(d)]. However, for larger optical thicknesses, the target position cannot be easily estimated from single-shot intensity images [Fig. S2(b)] due to the presence of scattered light. A ballistic contribution remains, as can be seen on the average image [Fig. S2(e)], but this contribution is significantly smaller than the random fluctuations caused by the scattered light [Fig. S2(f)]. In these conditions, the problem of detecting the exact position of the object becomes quite complicated, even in the presence of ballistic light. To illustrate this complexity, we have tested simple approaches based on maximal tracking and Gaussian fitting [Fig. S2(g) and Fig. S2(h)]; both of them fail at precisely estimating the target position at large optical thickness, while artificial neural networks are able to reach a much higher precision.

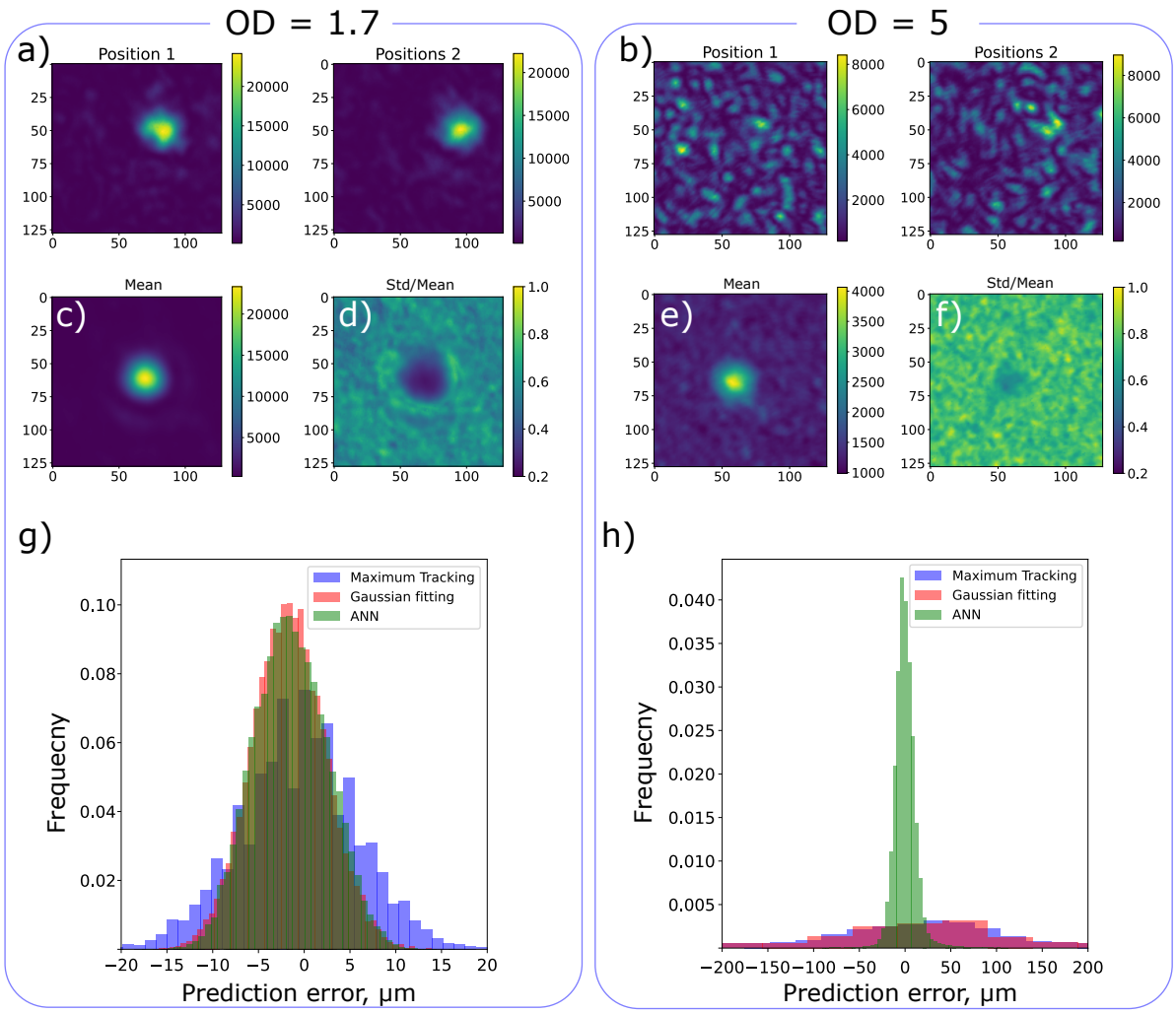

FIG. S2. (a),(b) Single shot intensity images extracted from the test set for two adjacent positions of the target, for optical thickness of (a) $b = 1.7$ and (b) $b = 5$. In the stronger scattering case, brighter spots are typically observed around the target position, but precisely localizing the target from such images is a difficult task. (c),(e) Average intensity calculated over 1000 frames, with the target located at the central position. (d),(f) Ratio between the standard deviation and the mean, which is equal to 1 for fully-developed speckles. A strong ballistic contribution can be seen in (d), as the ratio drops to 0.2. However, in (f), the minimal value is 0.85, which indicates that the contribution of the ballistic light is small as compared to the scattered one. (g),(h) Distribution of errors for different methods: maximum tracking, Gaussian fitting and ANN (CoordConv). In the strongly scattering case, the ANN performs much better than any of the simpler methods. In all images [(a) to (f)], axes are in camera pixels (6.5 µm).

S3. ESTIMATION OF THE FISHER INFORMATION

S3.1. Elimination of statistical dependence in-between pixels

Due to the finite extent of speckle grains, values measured by neighboring pixels are correlated to each other. While Eq. (2) of the manuscript intrinsically includes the effect of these correlations, it is in practice much more robust to first remove the dependencies by performing an appropriate change of basis.

Among those linear transformations that meet the task of decorrelating the random variable, we identify two particularly practical types of transformations. First, similar to the technique of principal component analysis [1], we can estimate the covariance matrix of the random variable from the dataset and then transform the basis such

that the covariance matrix becomes diagonal. By doing so, the lowest order dependencies (i.e. those captured by the covariance matrix) between different components of the transformed random variable are minimized. However, there could be higher order dependencies between these components, which would not be captured by the covariance matrix. While diagonalizing the covariance matrix is sufficient in many applications, it is possible to take these higher order dependencies into account even when restricting our analysis to linear transformations. The independent component analysis (ICA) [2] allows us to construct a linear transformation that minimizes the dependencies between components of the random variable. Unlike the technique used in principal component analysis, ICA harnesses more degrees of freedom when choosing the transformation matrix since there are fewer restrictions in the choice of the matrix (such as those related to the symmetry of the matrix).

In general, when dealing with data from an unknown distribution, one needs to first test whether the data follow Gaussian statistics. Whenever the data are Gaussian, ICA is not applicable. In such a case, however, the simpler alternative to the ICA, that is estimating the covariance matrix and decorrelating the random variable by rotating the basis such that the covariance matrix is diagonal, is sufficient to guarantee statistical independence between the transformed random variables. For non-Gaussian distributions, dependencies between components cannot be completely removed even by diagonalizing the covariance matrix; the ICA must be chosen in such cases.

In our experiments, we observe that the distributions are clearly non-Gaussian, hence we choose the ICA as a means of decorrelation. We thus apply an ICA in order to find the transition matrix A transforming a random vector X into its counterpart $Y = AX$ with approximately independent components. ICA requires that most of the components of Y do not follow a Gaussian distribution, and a suitable measure of non-Gaussianity is maximized in the process. We employ a computationally efficient version of ICA, the FastICA algorithm as described in [2]. Here, after whitening the random vector's components to unit variance, non-Gaussianity is maximized iteratively until convergence is reached. As a result of ICA, the probability density function of the transformed random vector Y approximately factorizes:

$$p(Y;\theta) \simeq \prod_{k=1}^{N} p_k(Y_k;\theta) \ , \tag{S1}$$

which allows us to employ the following simplified expression for the Fisher information matrix:

$$[\mathcal{J}(\theta)]_{ij} \simeq \sum_{k=1}^{N} \left\langle \left[\frac{\partial \ln p_k(Y_k;\theta)}{\partial \theta_i}\right] \left[\frac{\partial \ln p_k(Y_k;\theta)}{\partial \theta_j}\right] \right\rangle \ , \tag{S2}$$

where $p_k(Y_k;\theta)$ are the marginal distributions of the individual independent components.

S3.2. Mutual information analysis of the independent components

Performing an independent component analysis (ICA) does not guarantee that the components Y_i of the transformed random variable $Y = AX$ are independently distributed. Indeed, any given distribution is not necessarily related to a distribution with independent components by a linear transformation. In fact, since the speckle patterns follow complicated and unknown statistics, we do not expect the distribution $p(Y|\theta)$ to factorize exactly but only approximately. Modeling the distribution by a product requires us to employ a measure of dependence between components of Y. This is achieved by the concept of mutual information (MI) [3] between pairs of components of Y, which is defined as follows:

$$I(Y_i, Y_j) = \int dX \int dY p(Y_i, Y_j) \log\left(\frac{p(Y_i, Y_j)}{p(Y_i)p(Y_j)}\right) \ , \tag{S3}$$

where $p(Y_i, Y_j)$ is the joint distribution of the components Y_i and Y_j and $p(Y_i)$ and $p(Y_j)$ are their marginal distributions, respectively. Ideally, the MI between all distinct components vanishes since this corresponds to independently distributed components. We relax this condition by requiring that the ratio

$$g_{i,j} = \frac{I(Y_i, Y_j)}{I(Y_i, Y_i)} \tag{S4}$$

is small for all pairs of components (Y_i, Y_j), where the denominator coincides with the entropy of the random variable Y_i. $g_{i,j}$ measures the MI of a pair of pixels, relative to the entropy of one of the pixels. Moreover, we expect that the sum of this quantity over all components $\sum_j g_{i,j}$ is not too large. The MI is estimated using histograms (1-dimensional

histograms for the marginal distributions and 2-dimensional histograms for the joint distributions). For each fit, we use 10 bins of equal size. For each data set we use 50 000 data points to construct the linear transformation by the ICA and reserve 75 000 for the estimation of the MI. The worst case result of the sum $\sum_j g_{i,j}$, i.e., the largest value among all components Y_i is $\max(\sum_j g_{i,j}) \simeq (35.3, 3.2, 3.0, 2.6, 1.0, 0.5)$ for the Fisher information estimates shown in Fig. 3 of the manuscript, with the optical thicknesses $b = (0, 1.7, 2.5, 3.3, 4.2, 5)$ in this order. We observe that in the case of $b = 0$, some components of the random variable contain a considerable amount of information about the other components. Here, modeling the data with a product of marginal distributions should be considered as a rough approximation.

S3.3. Derivation of the Fisher information estimator of the marginal distributions

Applying Eq. (S2) entails estimating the probability density functions $p_k(Y_k;\theta)$ from a finite number of disorder configurations. This can be achieved by using histograms as approximate representations of $p_k(Y_k;\theta)$. In contrast to the conventional approach in density estimation where all bins are equal in width, using bins with equal frequency (i.e., equal fillings) yields a more stable estimate of the Fisher information, because small variations of the frequency do not influence the result.

Then, in order to approximate the probability density functions $p_k(Y_k;\theta)$ for different values of θ close to $\theta = (0,0)$, we take advantage of the transverse spatial invariance of the problem instead of physically moving the target. This allows us to estimate the partial derivatives of $p_k(Y_k;\theta)$ with respect to the parameters of interest with a centered finite difference scheme

$$\frac{\partial p_k(Y_k;\theta)}{\partial \theta_i} \simeq \frac{p_k(Y_k;\theta + \hat{e}_i\Delta\theta) - p_k(Y_k;\theta - \hat{e}_i\Delta\theta)}{2\Delta\theta}\,, \tag{S5}$$

where $\hat{e}_i$ is the i-th unit vector in the parameter space (this space is of dimension 2 in our case). To apply this procedure in practice, we first resize the images from 128×128 pixels to 32×32 pixels (which is the same resolution used for training the ANNs), and we then shift the images by $\Delta\theta$. Note that this approach results in a greater numerical stability as compared to resizing the images after the shift. We attribute this behavior to the fact that the resizing involves averaging over neighboring pixels; statistical fluctuations have a smaller influence on the resized pixels, which simplifies the interpolation between values of consecutive pixels that occurs when the images are shifted.

Our next task is to construct an estimator for the Fisher information of each marginal distribution $p_k(Y_k;\theta)$ using the finite difference approximation, based on the statistical sample that is available. Using statistical samples of the k-th marginal distribution $p_k(Y_k;\theta)$ as well as of the distributions $p_k(Y_k;\theta \pm \hat{e}_i\Delta\theta)$, we can estimate the Fisher information by constructing histograms for the distribution $p_k(Y_k;\theta)$, choosing the size of the bins such that the filling of the j-th bin $h_j^k = h^k$ is the same for each bin i. We denote the fillings of $p_k(Y_k;\theta \pm \hat{e}_i\Delta\theta)$ using the notation $h_j^{\pm,k}$, respectively. We assume that all samples are of the same size, which implies that $\sum_j h_j^k = \sum_j h_j^{\pm,k} = H$. With the width Δx_j of each bin that is determined by the filling h_j^k, we can approximate the probability density functions (PDFs) at the center of each bin by $p_j^k \simeq h_j/(H\Delta x_j)$ and $p_j^{\pm,k} \simeq h_j^{\pm,k}/(H\Delta x_j)$. The Fisher information with respect to the parameter θ can be written in a convenient way by

$$\mathcal{J}_{ii}^k(\theta) = \int dY_k p_k(Y_k;\theta)\left(\frac{\partial \ln p_k(Y_k;\theta)}{\partial\theta_i}\right)^2 = 4\int dY_k\left(\frac{\partial\sqrt{p_k(Y_k;\theta)}}{\partial\theta_i}\right)^2\,. \tag{S6}$$

Plugging the discretized estimate of the PDF into this equation and approximating the derivative by a symmetric finite difference yields

$$\mathcal{J}_{ii}^k(\theta) \simeq 4\sum_j\left(\frac{\Delta x_j}{2\Delta\theta}\left(\sqrt{p_j^{+,k}} - \sqrt{p_j^{+,k}}\right)\right)^2 = 4\sum_j\left(\frac{\Delta x_j}{2\Delta\theta}\left(\sqrt{\frac{h_j^{+,k}}{\Delta x_j H}} - \sqrt{\frac{h_j^{-,k}}{\Delta x_j H}}\right)\right)^2\,. \tag{S7}$$

Based on this expression, we can thus identify the following estimator for the Fisher information of the marginal distribution $p(Y_k;\theta)$:

$$\widehat{\mathcal{J}}_{ii}^k = \frac{1}{H(\Delta\theta)^2}\sum_j\left(\left(\sqrt{h_j^{+,k}} - \sqrt{h_j^{-,k}}\right)\right)^2\,. \tag{S8}$$

The total Fisher information is estimated by summing all contributions from the marginals using $\widehat{\mathcal{J}}_{ii} = \sum_k \widehat{\mathcal{J}}_{ii}^k$. In Eq. (S8), we recall that the index i denotes either the x or the y position of the target; this index enters the equations above only via the direction in which the images are shifted for the finite difference approximation. In the same way as for Eq. (S6), the off-diagonal elements of the Fisher information matrix are expressed by

$$\mathcal{J}_{ij}^k(\theta) = \int dY_k p_k(Y_k;\theta) \frac{\partial \ln p_k(Y_k;\theta)}{\partial \theta_i} \frac{\partial \ln p_k(Y_k;\theta)}{\partial \theta_j} \ , \tag{S9}$$

where $i \neq j$. The resulting estimator for these off-diagonal elements is

$$\widehat{\mathcal{J}}_{ij}^k = \frac{1}{H(\Delta\theta)^2} \sum_l \left(\sqrt{h_{il}^{+,k}} - \sqrt{h_{il}^{-,k}} \right) \left(\sqrt{h_{jl}^{+,k}} - \sqrt{h_{jl}^{-,k}} \right) \ , \tag{S10}$$

where the indices i and j in $h_{il}^{\pm,k}$ and $h_{jl}^{\pm,k}$ now denoting that the images where shifted in the direction corresponding to i and j, respectively. In practice, in our experiments, we observed that these off-diagonal elements are negligible compared to the diagonal elements, and we thus treat the problem as being effectively a single parameter estimation problem.

S3.4. Selection of the step size

Due to statistical fluctuations affecting histogram bin fillings, the approximation of the derivative and thus the estimate of the Fisher information depend on the value of the step size $\Delta\theta$. These fluctuations have a strong impact if the step size is too small. In contrast, if the step size is too large, we expect the finite difference approximation to be less accurate. Due to this trade-off, the optimal choice for $\Delta\theta$ is the smallest possible value for which the finite difference estimate of the Fisher information is not dominated by noise. We assume that this is the case when the estimate is stable with respect to small variations of $\Delta\theta$. To find this optimal value, we calculate the estimated Fisher information as a function of $\Delta\theta$ and observe that the estimated Fisher information is large for small $\Delta\theta$ and falls off rapidly when $\Delta\theta$ is increased. For larger $\Delta\theta$, the curves become almost independent on the step size and reach a plateau. We expect that the optimal step size is in the latter region since the finite difference approximation must not strongly depend on the choice of the step size. While one can easily find a reasonable step size by eye (by selecting one value where the curve flattens out), we employ here a formal criterion based on the second derivative of the curve, which allows us to identify the plateau as the region for which the slope remains constant. The second derivative fluctuates strongly for small $\Delta\theta$, while these fluctuations become small for larger $\Delta\theta$ when the plateau is reached. We then calculate the standard deviations of these fluctuations observed for large step sizes (when the plateau is already reached) by selecting a small window around a given step size. Our final selection of the optimal step size is the one where this standard deviation becomes greater than 10 times the standard deviation for the largest $\Delta\theta$ in the curves. In practice, this criterion corresponds to the selection of the smallest step size that belongs to the plateau.

In Fig. S3, we show the estimated Fisher information for the experimental datasets with different optical thicknesses. We observe that the curves show some features that are more complex as compared to numerically generated data [Fig. 2(b) and Fig. 2(c) of the manuscript]. However, they follow the same trend: a rapid decrease of the estimated Fisher information for small values of $\Delta\theta$, and a slow decrease of the estimated Fisher information for larger $\Delta\theta$. For cases in which a (tilted) plateau can be observed, we use the smallest step size where the second derivative of the curve vanishes, which corresponds to the onset of the plateau. If the second derivative does not vanish anywhere (as observed for an optical thickness of $b = 0$), we employ the same threshold as for the numerically generated data to identify the onset of the plateau. Finally, when oscillations are observed, we choose the location of the first local minimum to approximate the location of the onset of the plateau.

S3.5. Uncertainty in the estimated Cramér-Rao bound

Our method to estimate the Cramér-Rao bound from experimental measurements is inherently subject to several sources of errors.

A significant source of error comes from the approximation of the derivatives by a finite difference scheme and the associated selection of the optimal step size. In Fig. S3, we show the estimated Fisher information as a function of the step size in our experiments, for different optical thicknesses. From these curves, we can get an estimate of the error in the estimated Cramér-Rao bound by considering the maximum fluctuations of the curve in the region between the

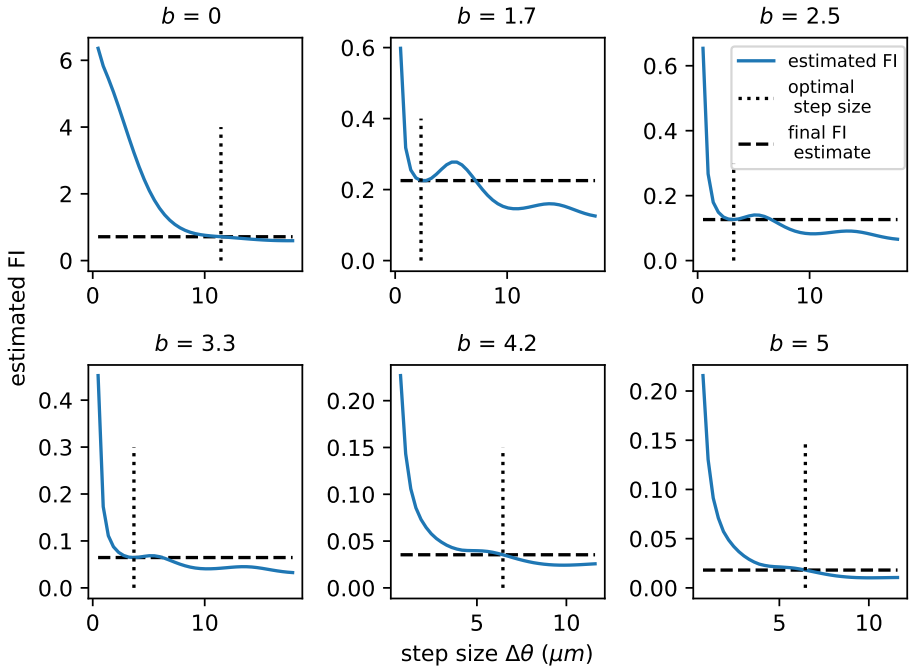

FIG. S3. The blue curves depict the estimated Fisher information for the experimental datasets as a function of the step size $\Delta\theta$ for different optical thicknesses b. The Fisher information is estimated using ICA for decorrelation and histograms to find the marginal densities. The vertical dotted lines correspond to step sizes where the Fisher information estimate shows good stability.

first and the second minimum (considering larger step sizes as being too large for the finite difference approximation to hold). In practice, we calculate the difference between the largest and the smallest Cramér-Rao bound in this region of step sizes, and we divide it by our estimated Cramér-Rao bound to obtain a relative error. The resulting error is typically around 30 %, but with a value of around 10 % for the optical thickness $b = 0$ and a value of around 40 % for the optical thickness $b = 5$. Note that, while this uncertainty in the absolute value of the Cramér-Rao bound is significant, the uncertainty in the relative values of the bound calculated from similar datasets is much lower (of the order of a few percents).

Another possible source of error is due to the finite number of samples used to estimate the probability densities. However, the size of our datasets is sufficient so that we can neglect this source of error, both in simulations and in experiments. In simulations, we can verify this by repeating the whole analysis 100 times with different noise realizations, and calculate the mean μ_{FI} and the standard deviation σ_{FI} of the estimated Fisher information. We find $\mu_{\mathrm{FI}}/\mathcal{J} = 1.01$ and $\sigma_{\mathrm{FI}}/\mathcal{J} = 0.05$ for the Gaussian distribution and $\mu_{\mathrm{FI}}/\mathcal{J} = 0.97$ and $\sigma_{\mathrm{FI}}/\mathcal{J} = 0.03$ for the non-Gaussian distribution. In the experiments, in order to check the influence of the size of the dataset, we used only a fraction of the original datasets and verify that the estimated Fisher information remains constant. We observed that our choice of the optimal step size and the corresponding Fisher information estimate does not change with smaller sample sizes, unless the sample size becomes very small. As an example, we show in Fig. S4 the procedure of estimating the Fisher information by selecting the optimal step size for a single dataset ($b = 4.2$) but for different sizes of the dataset. We observe that, taking either the full dataset (with a size of 125000) or only half of it, we obtain nearly the same location for the first plateau of the curve curve, and thus nearly the same estimated Fisher information. When using only a fourth of the full dataset, the estimated Fisher information decreases slightly but still yields a reasonable value, the plateau being less pronounced in this case.

Finally, note that another source of error could be due to the fact that the ICA does not perfectly decorrelate the data. However, this source of error is unlikely to significantly affect our estimates of the Fisher information. Indeed, we have observed in simulations that the true Fisher information is reached using our approach even for correlated data (see Fig. 2 of the manuscript). Moreover, in the experiments, we observed that the mutual information of

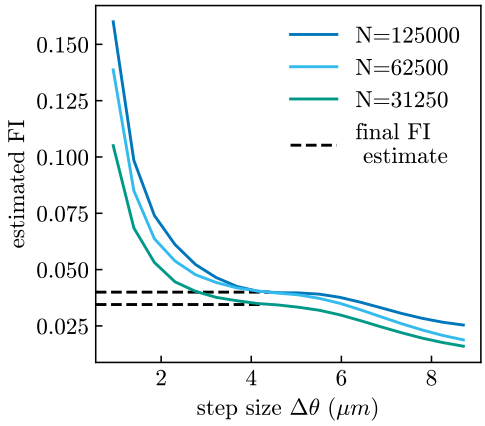

FIG. S4. The estimation of the Fisher information by selecting the optimal step size is shown for different sample sizes N. The curves depict the Fisher information as a function of the step size $\Delta\theta$ for one example dataset (optical thickness $b = 4.2$). The horizontal dashed lines indicates the final estimates of the Fisher information where the curves show the first plateau.

the transformed variables is small (see Section S3.2), which indicates that the ICA effectively managed to remove statistical dependence in-between the components of the transformed variables.

S4. ANN ARCHITECTURE

S4.1. Data processing procedure

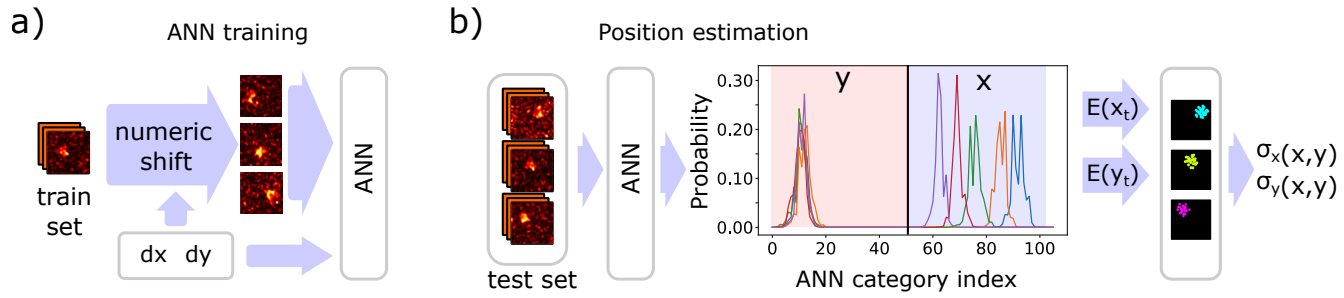

FIG. S5. (a) ANNs are trained using data augmentation by numerically shifting data measured with the target located at the central position. (b) ANNs predict the x and y coordinates of the target as a probability density, thanks to the softmax activation function used in the last layer. As an example, we show the probability densities given by the ANN for target positions in the top row ($b = 4.2$ dataset). In this example, the x coordinate of the target changes, while the y coordinate remains fixed. The left 53 category indices correspond to the y coordinate estimates, and the right 53 category indices correspond to the x coordinate estimates. The actual target position is inferred as the expected values of those distributions. The achieved precision is calculated as the standard deviation of the statistics of the predicted position for different test examples.

S4.2. Hyperparameter tuning

As the accuracy and the precision of the predictions may depend on the network architecture, we had to tune the ANN's hyperparameters to achieve optimal performance. The layouts of the architectures we tested (Dense, Convolutional, Convolutional with coordinate layers, and Dense Convolutional) are shown in Fig. S6. The parameters

to be tuned for each architecture and for $b = 4.2$ are outlined in Tables S1 to S3. The dependence of the optimal CoordConv hyperparameters on the optical thickness is outlined in Table S4.

For each of the networks, the "Depth" parameter denotes the number of layers, and the "Out size" parameter specifies the discretization of the predicted position probability distribution. The detailed description of the DCCNN hyperparameters can be found in Ref. [4]. The parameters were optimized using a grid search: for each combination, the corresponding ANN has been trained for 3 epochs and the values of loss, validation loss and prediction MSE were recorded. Using this strategy, we have observed that the CoordConv architecture performs best for a number of trainable parameters that is approximately constant over all optical thicknesses (around 5×10^5 parameters, while the number of trainable parameters was swept from 1.6×10^5 to 96.7×10^5 for each optical thickness). However, the automatic hyperparameter tuning strategy is based on 3 epochs, which might not be representative of the behavior of the ANNs at the end of the training; it thus remains plausible that, with a larger number of parameters, the network could perform better at high optical thicknesses.

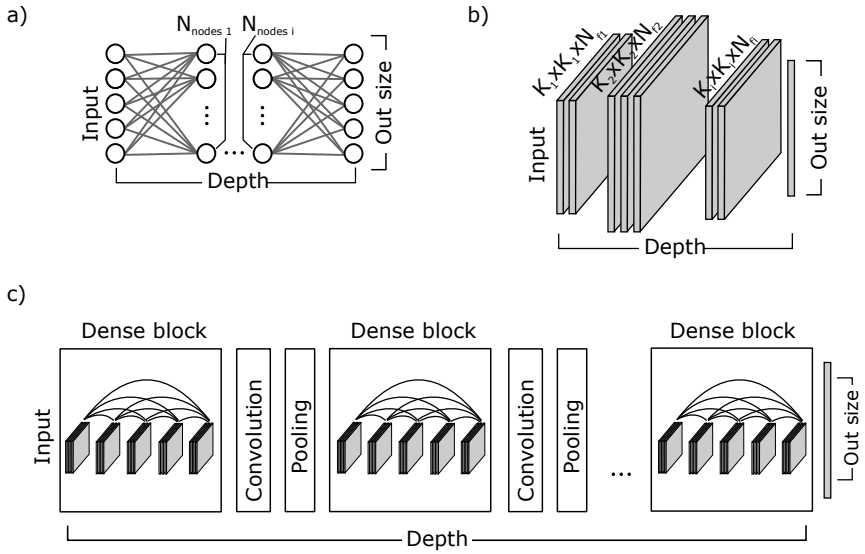

FIG. S6. ANN layouts. (a) Dense (DNN). (b) Convolutional (CNN), for the Convolutional with coordinate layers (CoordConv) the two coordinate layers (rescaled -1 to 1) were added after each layer stack. (c) Densely Connected Convolutional (DCCNN).

Parameter	Tuning range	Optimal value
Depth	2-12	10
Out size	13-43	13
Number of nodes in layers 1-2	20-820	420
Number of nodes in layer 2-10	20-820	420
Trainable parameters $\times 10^5$	4.2-86.9	8.7

TABLE S1. Tuning ranges and optimal parameters of the DNN.

Parameter	Tuning range	Optimal value CNN	Optimal value CoordConv
Depth	2-5	4	2
Out size	13-63	53	58
Filter shape	6×6 - 10×10	7×7	8×8
Number of filters for each layer	20-100	90	20
Trainable parameters $\times 10^5$	1.6-96.7	3.3	5.5

TABLE S2. Tuning ranges and optimal parameters of the CNN and CoordConv

Parameter	Tuning range	Optimal value
Depth	2-6	5
Out size	13-63	33
Dense block depth	2 - 6	2
Dense block growth	20-60	30
Dense block bypass	20-60	30
Trainable parameters $\times 10^5$	3.4-240	35

TABLE S3. Tuning ranges and optimal parameters of the DCCNN.

Parameter	none	b=1.7	b=2.5	b=3.3	b=4.2	b=5
Depth	3	4	3	3	2	2
Out size	53	58	13	48	58	23
Filter shape	7×7	6×6	6×6	6×6	8×8	7×7
Number of filters for each layer	20	20	30	20	20	30
Trainable parameters $\times 10^5$	5	5.1	3.1	4.3	5.5	4.3

TABLE S4. Optimal CoordConv parameters for different optical thicknesses. For all optical thicknesses, the tuning range used for the hyperparameter search is the same as in Table S2.

S4.3. Selection of the best architecture

We could select the optimal architecture as the one that minimizes either training loss, validation loss, or prediction MSE. In order to compare the ANN architectures, we considered the $b = 4.2$ dataset and trained them for 50 epochs. We then compared the standard deviation of the position estimates, the validation loss and the MSE, which are plotted in Fig. S7. In addition to the architectures mentioned in the manuscript, we also tested here a Vision Transformer (VT) ANN [5] with 8 transformer layers, each of these layers including a multi-head self-attention mechanism with 4 parallel attention heads, a feed-forward network with 128 hidden layer units and a GELU activation function, and normalization layer. The input images are divided into 4×4 patches, which leads to the feature space of a dimension of 64. The target coordinates are encoded using the same 2-hot encoding scheme as in the other architectures, with each coordinate represented as a 64-dimensional vector.

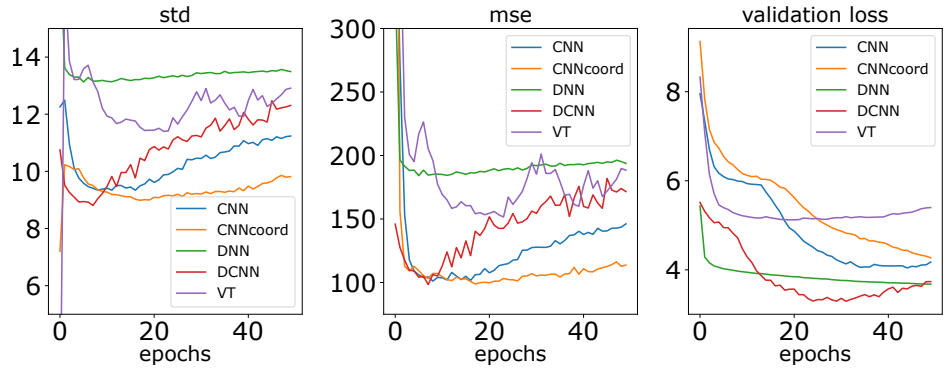

FIG. S7. Performance comparison of different ANN architectures for the $b = 4.2$ dataset. Left panel: standard deviation of the predicted data averaged over all x and y positions. Central panel: mean squared error (with respect to the ground truth). Right panel: validation loss.

As can be seen from Fig. S7, the CNN and the CoordConv architectures can reach roughly the same standard deviation, but CoordConv reaches a slightly smaller MSE due to a lower bias. Therefore, we consider it as being the optimal architecture for the task. However, we need to mention that we did not perform hyperparameter optimization with VT in the same way as we did for other architectures, as the search space is much larger than for the other architectures; therefore, we cannot exclude that better performances could be obtained by finding better hyperparameters.

S4.4. Bias correction

As can be seen from Fig. 3 of the manuscript, all ANN architectures develop a bias. To compare the ANN predictions to the unbiased Cramér-Rao bound, we need to correct for this bias. For this purpose, we construct a function that characterizes the dependence of the bias on the target position. Initially, we extract a set of patterns from the test dataset, and apply the same augmentation procedure as employed during the training phase. We then pass this set through the trained network and calculate the average difference between the predicted and real values. After that, we fit a 2D spline function $B(x, y)$ to these values, and further use it to correct the predicted coordinates. An example of the fitted $B(x, y)$ (for x and y coordinates) and the effect of the bias correction on the ANN predictions is shown in Fig. S8.

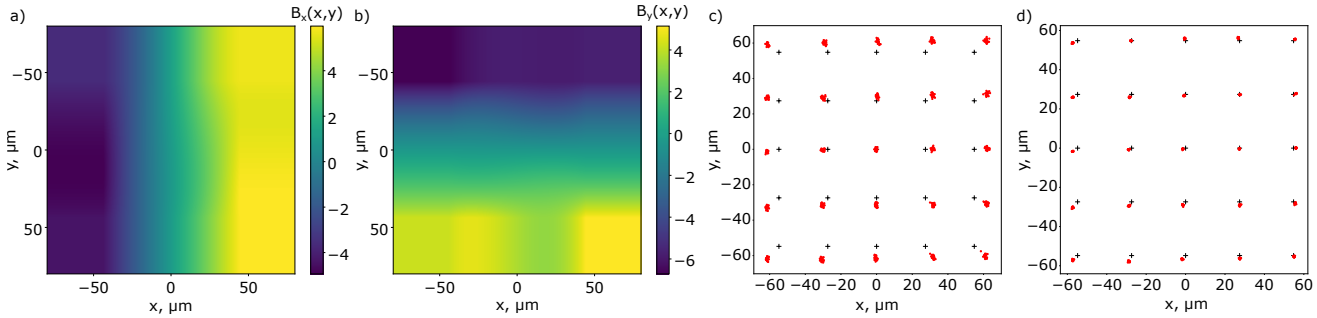

FIG. S8. (a),(b) Bias correction functions in x and y. (c) Predicted target position values (red dots) for $b = 2.5$ before and (d) after bias correction. The black crosses in (c) and (d) denote true positions.

S4.5. Influence of the size of the test set

From Fig. 3(b) of the manuscript, it seems that the Cramér-Rao bound can sometimes be overpassed, for some ANN realizations and some target positions. We attribute this effect to the finite sample size of the test set. In Fig. S9, we plot the ANN prediction histograms while increasing the size of the test set: we first use a quarter of the

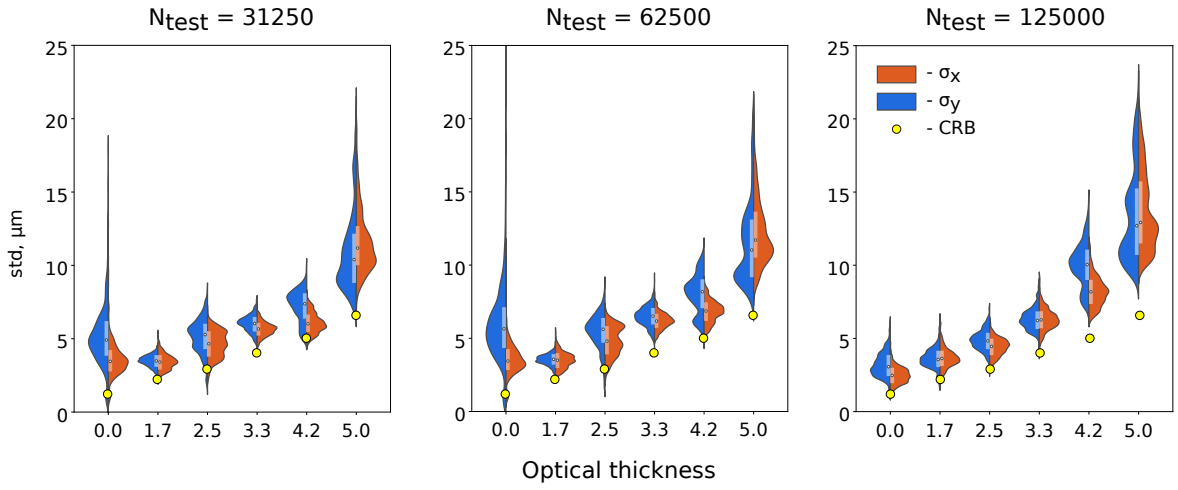

FIG. S9. Violin plots showing the CoordConv precision achieved for 25 random initializations of the network, for different sizes of the test set. White dots show the median values of σ_x and σ_y, vertical bars indicate the corresponding first and third quartiles, and the colored areas show the associated histograms. The yellow circles represent the Cramér-Rao bound (CRB).

test data, a half, and finally the full test set. As can be seen, increasing the size of the test set reduces the number of configurations that overpass the bound, which indicates that the finite sample size of the test set is likely to explain the few occurrences in which the Cramér-Rao bound seems to be overpassed.

S4.6. Target position dependence

In Fig. S10, we plot the dependence of the average (over 25 initializations, as well as over x and y coordinates) ANN uncertainty on the target position. As can be seen from this plot, the average uncertainty does not depend on the target position for the weakly-scattering samples. When the optical thickness is increased, however, the reduction of the ANN precision does not happen uniformly over all target positions. We indeed observe that the central positions are easier to predict for the network, as compared to the ones located in the corners of the field of view. We attribute this effect to the fact that the corner positions are less connected to the output of the network, which makes it harder for the information to flow from these areas to the output.

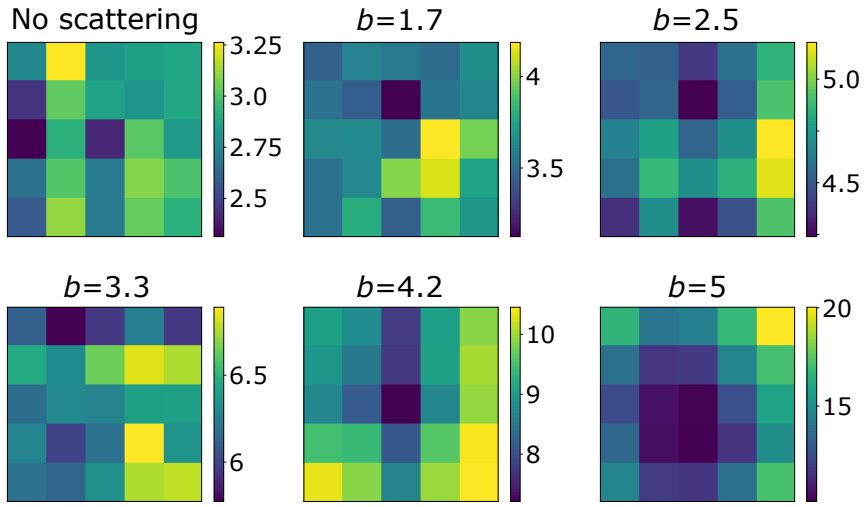

FIG. S10. Dependence of the ANN uncertainty (average for x and y) on the target position averaged over 25 ANN random initializations.

S5. MODEL GENERALIZABILITY

S5.1. Influence of the scattering strength

ANNs can struggle to generalize to conditions that are not represented in the training dataset. In our experiments, we trained multiple models using experimental data measured for different scattering strengths. It is interesting to assess whether these ANNs can be used to precisely estimate the target position for unseen scattering densities. In Fig. S11, we present cross-validation results, where models trained at a specific optical thickness are tested on data from different densities. We observe that models trained on stronger scattering datasets perform relatively well when tested in weakly scattering conditions, while models trained on weakly scattering datasets becomes much more imprecise when tested in strongly scattering conditions. In order to generalize, it is thus preferable to train the model in the worst-case scenario with regards to the scattering strength.

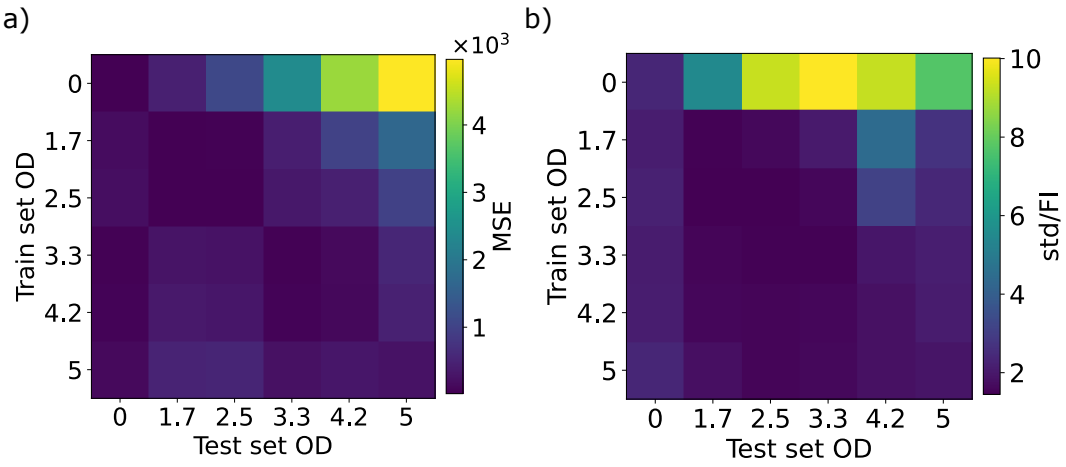

FIG. S11. (a) Mean squared error and (b) standard deviation normalized by the value of the Cramér-Rao bound of the models trained on one optical thickness and tested on another.

S5.2. Influence of the object size and shape

The size of the object has an influence on the localization accuracy. For an object with sharp edges, as in our experiments, information about the target position is only created at the edges of the object [6]. In particular, for a square object of side length a, in an ideal free-space imaging configuration and for Gaussian noise statistics, the Fisher information would scale with a and the Cramér-Rao bound would scale with $1/\sqrt{a}$. To verify if a similar behavior can be observed in our experiments (in the presence of a scattering medium), we considered a scattering medium of optical thickness $b = 2.5$ and we performed measurements on three different objects of different sizes (2×2 DMD pixels, 5×5 DMD pixels and 7×7 DMD pixels). From these measurements, we calculated a Cramér-Rao bound of 2.9 µm for the biggest object, 3.1 µm for the medium-sized object and 5.5 µm for the smallest object. This shows that a bigger object indeed produces a larger amount of information, and therefore yields a smaller Cramér-Rao bound. However, the scaling in $1/\sqrt{a}$ is only approximately recovered, indicating that other effects due to the presence of the scattering medium could play a role (such as the complex noise statistics or possible diffraction effects).

As reported in the manuscript, the observed precision of the ANN is, on average, always higher than the Cramér-Rao bound. Here, we observed a value of 5.2 µm for the biggest object, 5.5 µm for the medium-sized object and 20.7 µm for the smallest object. While the ANN precision stays relatively close to the bound for the biggest object and for the medium-sized object, it becomes significantly higher than the bound for the smallest object. This confirms that, for less favorable signal-to-noise conditions (either due to a smaller object size or to a larger optical thickness), the bound becomes harder to reach.

We also perform similar measurements on a cross-shaped object, covering 8×8 DMD pixels with a line thickness of 3 pixels. We obtained a Cramér-Rao bound of 3.0 µm and an ANN precision of 5.4 µm, very close to what is observed for square objects of similar size.

Finally, in addition to the cross-test performed for different scattering strengths (see Fig. S11), we performed a cross-test on different object sizes and shapes to evaluate the ability of the ANN to generalize. The results are shown in Fig. S12. For objects with comparable amount of information (1st, 2nd and 4th rows and columns), the ANN is able to generalize relatively well when trained with a different object sizes and shapes, with an increase of the standard deviation ranging from 13% to 78% when the ANN is trained and tested with a different object size or shape (depending on the geometrical similarities between the considered objects). In all cases involving the smallest object (3rd raw and column), the situation changes as the amount of information in the testing data is significantly smaller. Then, training the ANN on different object sizes and shapes leads to poor results (3rd column). Note that, when trained on the smallest object and evaluated using images with more information (3rd line), the standard deviation decreases: the larger amount of information available in the data then compensates the penalty induced by training and testing on a different object size and shape.

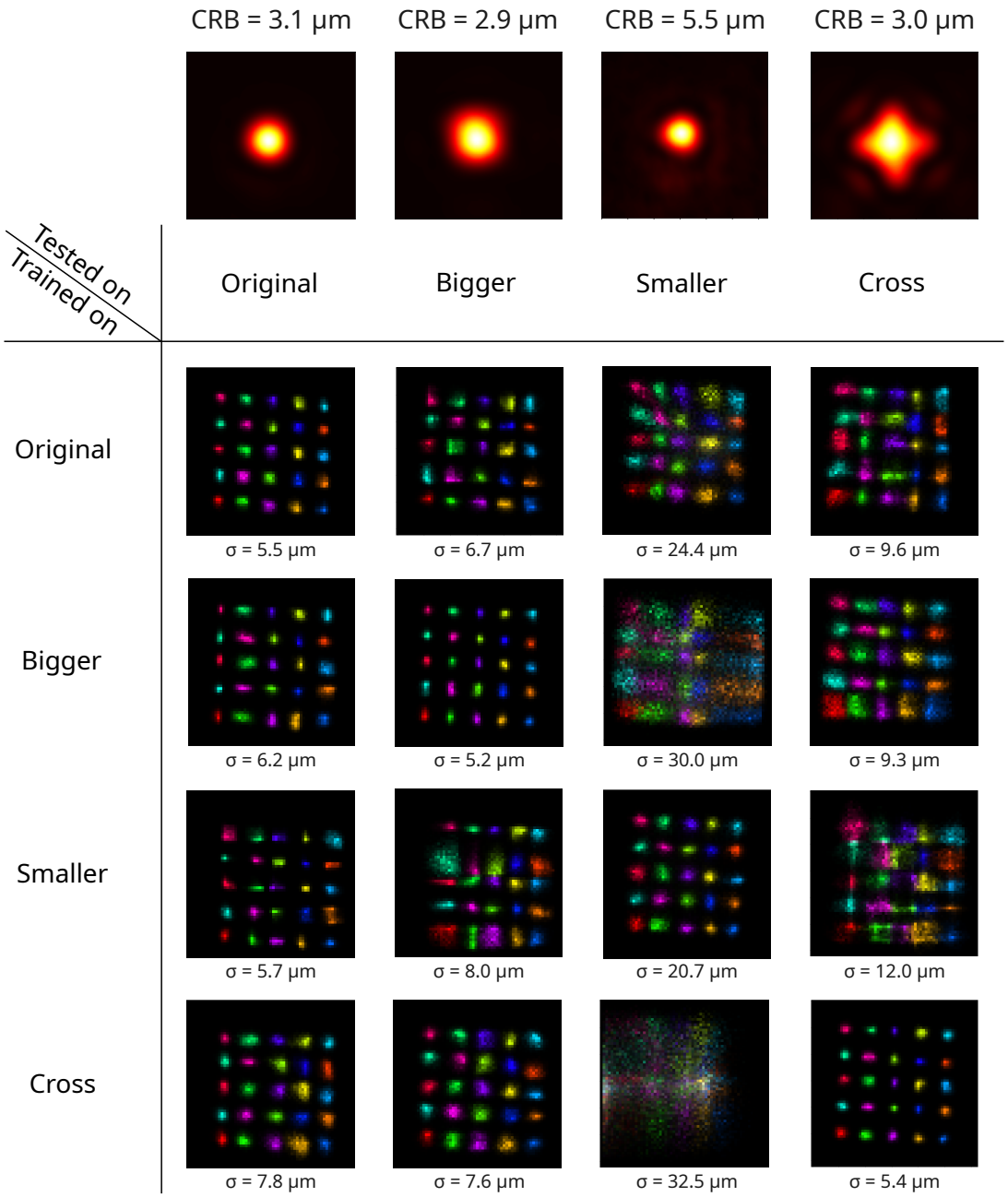

FIG. S12. Standard deviation σ of the models trained on one object size and tested on another. The original object is the one presented in the manuscript (composed of 5×5 DMD pixels), the bigger object is composed of 7×7 DMD pixels, the smaller object is composed of 2×2 DMD pixels, and the cross covers 8×8 DMD pixels (with a line thickness of 3 pixels). Training and testing are performed with an optical thickness $b = 2.5$. Images of the objects (top row) correspond to the ballistic contributions that are obtained by averaging over many disorder realizations; they are presented for illustration purposes. Note that, due to slight misalignment between the DMD and the camera, the ground truth positions appear as slightly tilted in this figure. This effect is consistent for the different targets and does not affect the ANN standard deviation estimates.

[1] I. T. Jolliffe and J. Cadima, Principal component analysis: a review and recent developments, Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences **374**, 20150202 (2016).
[2] A. Hyvärinen and E. Oja, Independent component analysis: algorithms and applications, Neural Networks **13**, 411 (2000).

[3] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 1999).
[4] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society, Los Alamitos, CA, USA, 2017) pp. 2261–2269.
[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in *International Conference on Learning Representations (ICLR)* (2021).
[6] J. Hüpfl, F. Russo, L. M. Rachbauer, D. Bouchet, J. Lu, U. Kuhl, and S. Rotter, Continuity equation for the flow of Fisher information in wave scattering, Nature Physics **20**, 1294 (2024).