Opto-Electronic Advances

Original Article

2021, Vol. 4, No. 5

DOI: 10.29026/oea.2021.200016

Deep-learning-based ciphertext-only attack on optical double random phase encryption

Meihua Liao[1†], Shanshan Zheng[2,3†], Shuixin Pan[1], Dajiang Lu[1], Wenqi He[1], Guohai Situ[2,3,4*] and Xiang Peng[1*]

Optical cryptanalysis is essential to the further investigation of more secure optical cryptosystems. Learning-based attack of optical encryption eliminates the need for the retrieval of random phase keys of optical encryption systems but it is limited for practical applications since it requires a large set of plaintext-ciphertext pairs for the cryptosystem to be attacked. Here, we propose a two-step deep learning strategy for ciphertext-only attack (COA) on the classical double random phase encryption (DRPE). Specifically, we construct a virtual DRPE system to gather the training data. Besides, we divide the inverse problem in COA into two more specific inverse problems and employ two deep neural networks (DNNs) to respectively learn the removal of speckle noise in the autocorrelation domain and the de-correlation operation to retrieve the plaintext image. With these two trained DNNs at hand, we show that the plaintext can be predicted in real-time from an unknown ciphertext alone. The proposed learning-based COA method dispenses with not only the retrieval of random phase keys but also the invasive data acquisition of plaintext-ciphertext pairs in the DPRE system. Numerical simulations and optical experiments demonstrate the feasibility and effectiveness of the proposed learning-based COA method.

Keywords: optical encryption; random phase encoding; ciphertext-only attack; deep learning

Liao MH, Zheng SS, Pan SX, Lu DJ, He WQ et al. Deep-learning-based ciphertext-only attack on optical double random phase encryption. *Opto-Electron Adv* **4**, 200016 (2021).

Introduction

Optical encryption has captured growing attentions in the past two decades owning to its inherent advantages such as parallel signal processing and high dimensional operation[1,2]. When irradiating a two-dimensional image vertically with parallel light, all points on the image will be modulated by light simultaneously. Besides, inherent parameters of the optical system (e.g., the amplitude, phase, wavelength, polarization and diffraction distance) can be designed as the security keys of an optical cryptosystem for multi-dimensional encryption[2]. Most of optical encryption techniques are based on the technique proposed by P. Refregier and B. Javidi, and are known as double random phase encryption (DRPE)[3], which uses two statistically independent random phase masks (RPMs) as the security keys to scramble an original plaintext image into a stationary white noise. Following this pioneering work, a variety of modifications of

[1]Key Laboratory of Optoelectronic Devices and System of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China; [2]Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China; [3]Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China; [4]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310000, China.

[†]These authors contributed equally to this work.

[*]Correspondence: GH Situ, E-mail: ghsitu@siom.ac.cn; X Peng, E-mail: xpeng@szu.edu.cn

Received: 18 June 2020; Accepted: 25 August 2020; Published: 20 May 2021

DRPE have been developed in other linear canonical transform domains such as the fractional Fourier domain[4,5], Fresnel domain[6] and wavelet domain[7]. In the meantime, random phase encoding has been incorporated with typical optical signal processing or imaging architectures, such as digital holography[8,9], joint transform correlator[10], interference[11], diffractive imaging[12], computational ghost imaging[13] and ptychography[14]. Note that the security performance of a cryptosystem is one of the major concerns. A cryptosystem can be considered sufficiently secure if and only if it can endure safety assessment by cryptanalysis. Cryptanalysis refers to the study on cryptosystems that aims to identify any defect in them that will permit retrieval of the plaintext from the ciphertext, without necessarily knowing the secret key[15]. On the other hand, a variety of attacks on existing optical cryptosystems can facilitate the development of security-enhanced optical cryptosystems[16–18]. As fueled by the advancement of optical cryptography, many optical cryptanalysis methods have been proposed as well, ranging from the chosen plaintext attacks (CPA)[19,20], known plaintext attacks (KPA)[21,22], and ciphertext-only attacks (COA)[23–25]. Among these optical cryptanalysis methods, CPA and KPA require attackers to access more resources and more control of the encryption system, and COA is usually considered as the most critical yet challenging problem since only a minimum resource is available to break the cryptosystem. In the existing optical cryptanalysis methods, the COA problem is usually transferred to a phase retrieval problem with single intensity measurement. The iterative phase retrieval algorithm[26–34] has been employed to solve this problem by exploiting an estimated signal domain support[26] and a given frequency domain constraint. However, it is time-consuming since it usually needs thousands of iterations to converge to a feasible solution. It has also been demonstrated that speckle correlation techniques are feasible for COA due to the high similarity between the autocorrelations of the ciphertext and the plaintext[35]. However, for the coherent DRPE system, the autocorrelation of ciphertext or the energy spectral density is usually contaminated by speckle noise produced by the RPM at the input plane, which is actually shown in a way similar to the "shower-curtain effect"[36]. One of the methods to remove this kind of speckle noise is spatial averaging, which is done by dividing a ciphertext image into a sequence of sub-images[37]. But there is a trade-off between the number and the size of sub-images, which limits the spatial resolution of the recovered image.

Over the past few years, deep learning has attracted increasing attentions and found to be highly flexible in solving various types of ill-posed inverse problems in optical sensing and imaging[38–53] such as optical tomography[41], computational ghost imaging[42,43], visual tracking[44], digital holography[45,46], lensless phase imaging[47–49], as well as imaging through scattering media[50–53]. For optical cryptanalysis, deep learning has demonstrated its capability of attacking on several optical cryptosystems such as DRPE and triple random phase encoding[54], interference/diffraction encryption[55–57] and computer-generated hologram encryption[58]. However, all of the aforementioned learning-based attack methods belong to the category of CPA, which requires a large set of plaintext-ciphertext pairs of a cryptosystem under analysis. This will be challenging because it is unlikely for an attacker to have access to the cryptosystem for such a long time.

Here, we demonstrate methodologically, numerically and experimentally for the first time, to our knowledge, that the use of deep learning can solve the inverse problems in COA against the classical DRPE. To be specific, we develop a two-step deep learning framework that retrieves the plaintext from an intercepted unknown ciphertext alone. For acquiring the training data, we construct a virtual DRPE system that includes different random phase keys to provide the statistically ergodic property of the speckle pattern. We note that the autocorrelation of the ciphertext in DRPE contains the information of the autocorrelation of the plaintext, only that the former one is with some additive speckle noise. Inspired by the principle of speckle correlations, we divide the inverse problem in COA into two inverse problems: one is the removal of the speckle noise from the autocorrelation of the ciphertext, and the other is the retrieval of the plaintext from the noise-free autocorrelation. Accordingly, two cascaded deep neural networks (DNNs) are employed to respectively solve the two specific inverse problems. With appropriate training, the two trained DNNs can be easily used to predict the plaintext image from the unknown ciphertext without knowing the phase keys.

Principle and method

Learning-based COA approach

In DRPE, a plaintext can be encrypted into a white noise-like distribution by employing an optical 4*f* system, where two RPMs serving as the keys are placed at the input plane and Fourier plane, respectively. The optical structure of DRPE is shown in Fig. 1(a). The encryption process can be mathematically expressed as

$$C(x,y) = \mathrm{FT}^{-1}\{\mathrm{FT}\{P(x,y)\cdot M(x,y)\}\cdot N(u,v)\} \ , \qquad (1)$$

where $\mathrm{FT}\{\cdot\}$ and $\mathrm{FT}^{-1}\{\cdot\}$ represent the Fourier transform and inverse Fourier transform, respectively; (x,y) and (u,v) represent the coordinates of the spatial domain and Fourier domain, $P(x,y)$ and $C(x,y)$ denote the plaintext and the ciphertext respectively; $M(x,y)=\exp[\mathrm{i}2\pi r_1(x,y)]$ and $N(u,v)=\exp[\mathrm{i}2\pi r_2(u,v)]$ represent the two RPMs, where $r_1(x,y)$ and $r_2(u,v)$ are the statistically independent uniform distribution in the range of $(0,1]$. The decryption procedure of DRPE is the exact inverse process of the encryption with the conjugates of phase keys.

The encryption process of the DRPE can be considered as the forward propagation process (see Fig .1(a)), and it is defined as $C=\boldsymbol{T}\{P\}$, where $\boldsymbol{T}\{\cdot\}$ denotes the forward model. Optical cryptanalysis is a typical inverse problem that is denoted as $P=\boldsymbol{T}^{-1}\{C\}$ (see Fig .1(b)), aiming to retrieve the plaintext from the corresponding ciphertext. One possible pure data-driven solution is to train a DNN with a set of plaintext-ciphertext pairs. Assuming that a training set of ground-truth plaintexts and their ciphertexts $\{P_i, C_i\}$ is known. The optical cryptanalysis problem implicitly solved by DNN can be formulated as the following equation

$$R_{\mathrm{DNN}} = \underset{\theta\in\Theta}{\arg\min}\,\|R_\theta\{C_i\}-P_i\|^2 \ , \qquad (2)$$

where R_θ is the mapping function of the DNN, and θ denotes the parameters of weights and biases. This kind of "end-to-end" mapping method is simple but needs a large number of plaintext-ciphertext pairs in the same optical encryption system, which means that an attacker has the ability to access the cryptosystem in advance.

Nevertheless, for COA, according to Kerchhoff's principle[15], the attacker is assumed to have access only to a ciphertext and use it alone to retrieve the corresponding plaintext. In practice, for the COA on DRPE, the attacker still has the knowledge of the cryptosystem (e.g., coherent illumination, 4*f* architecture) except for the random phase keys. Therefore, it is possible to gather a set of training data from a virtual DRPE system that includes a set of randomly generated RPMs placed at the spatial and frequency domain. Since the ciphertext is obtained from the different plaintext and different RPMs, the "end-to-end" mapping DNN carries the burden of learning all of the physical laws.

We have noted that the DRPE cryptosystem is essentially a coherent imaging system and the encryption formulation (Eq. (1)) can be rewritten as

$$C(x,y) = P'(x,y) * h(x,y) \ , \qquad (3)$$

where the symbol "$*$" denotes a convolution operation, $P'(x,y)=P(x,y)\cdot M(x,y)$ is the complex field at the input plane, $h(x,y)=\mathrm{FT}\{N(u,v)\}$ is the point spread function (PSF) of the DRPE system. Then the autocorrelation of ciphertext $Ac(x,y)$ can be written as

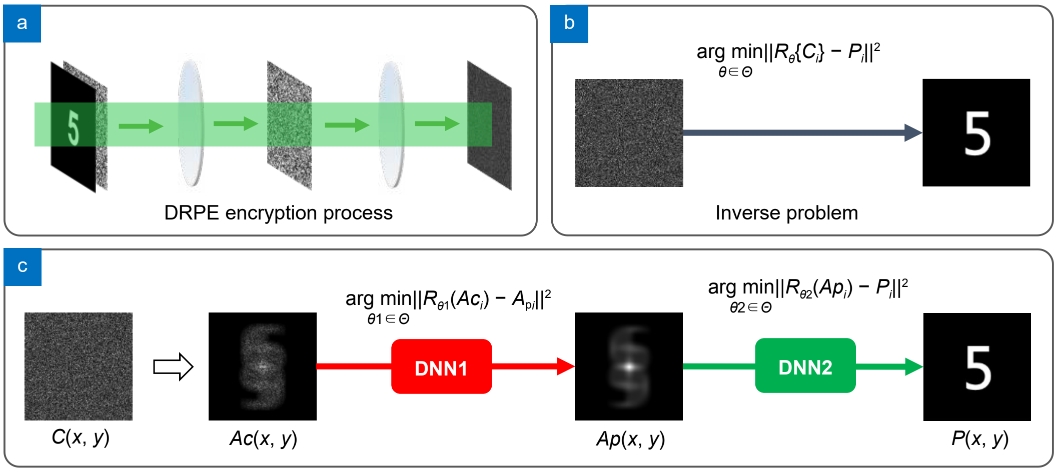
a
5
DRPE encryption process
b
arg min||R_θ{C_i} − P_i||^2 $\theta\in\Theta$
5
Inverse problem
c
arg min||$R_{\theta1}$(Ac_i) − A_{pi}||^2 $\theta1\in\Theta$
arg min||$R_{\theta2}$(Ap_i) − P_i||^2 $\theta2\in\Theta$
DNN1
DNN2
5
C(x, y)
Ac(x, y)
Ap(x, y)
P(x, y)

Fig. 1 | Overview of learning-based COA on DRPE. (**a**) The encryption process of DRPE is a forward propagation process. (**b**) The COA is an inverse problem, aiming to obtain an optimized estimate of the plaintext from the ciphertext. (**c**) Flowchart of the proposed COA method, where two DNNs (DNN1 and DNN2) are used in serial to respectively learn the removal of speckle noise from the autocorrelation of the ciphertext *Ac* and the prediction of the final plaintext *P* from its autocorrelation *Ap*.

$$\begin{aligned}Ac(x,y) =& C(x,y)\otimes C(x,y)\\ =&[P'(x,y)\otimes P'(x,y)]*[h(x,y)\otimes h(x,y)]\\ =&[P'(x,y)\otimes P'(x,y)]*\delta(x,y)\ ,\end{aligned}\tag{4}$$

where the symbol "⊗" denotes the autocorrelation operation, $\delta(x,y)$ is a peaked function. It suggests that the contribution of RPM2 at the Fourier plane can be removed by performing an autocorrelation operation. Therefore, the calculated autocorrelation of the ciphertext can be simplified as

$$Ac(x,y)\approx[P'(x,y)\otimes P'(x,y)]=Ap(x,y)+S(x,y)\ ,\tag{5}$$

where $Ap(x,y)$ denotes the autocorrelation of plaintext, $S(x,y)$ is the speckle noise term, which refers to the contribution of RPM1. If this speckle noise can be removed as well, the plaintext can be further retrieved. Decorrelation or reconstruction of an object from the modulus of its Fourier transform is a long-standing challenge and it is essentially an ill-posed inverse problem due to the absence of its Fourier phase. Usually, this type of inverse problem can be solved by an iterative phase retrieval algorithm with some prior knowledges such as non-negative and real-valued object and the support area in the object domain[26,27]. However, iterative methods are time-consuming, which makes real-time reconstruction a challenge. Recently, deep learning offers an alternative approach to perform such de-correlation tasks[48,51,52], and U-net shows powerful performance in solving image pixel regression problems[59]. Accordingly, a U-net neural network can be employed to perform the de-correlation task by training the mapping relationship between the autocorrelation pattern (which corresponds to the magnitudes of an object's Fourier transform) and the real-space object from a large prepared training set. After the training, the de-correlation DNN model can invert an autocorrelation pattern to the corresponding object image in real-time.

Therefore, the problem to be addressed in COA on DRPE can be reformulated as two inverse problems: one is the removal of the speckle noise from the autocorrelation of the ciphertext while the other is the retrieval of the plaintext from the noise-free autocorrelation. Inspired by the aforementioned analysis while aiming at achieving a better performance with limited training data, we propose a two-step deep learning strategy for solving the problems of the COA on DRPE (see Fig. 1(c)). Specifically, the autocorrelation functions of the ciphertext and the plaintext should be calculated first as the feature to be trained. Then two cascaded DNNs are built to solve the two corresponding inverse problems, DNN1 takes the autocorrelation of the ciphertext $Ac(x,y)$ as its input and estimates the plaintext autocorrelation $Ap(x,y)$; DNN2 is trained to predict the final plaintext image $P(x,y)$ from $Ap(x,y)$. Mathematically, these two inverse problems solved by DNN1 and DNN2 can be respectively formulated as

$$R_{\text{DNN1}}=\underset{\theta1\in\Theta}{\arg\min}\,\|R_{\theta1}(Ac_i)-Ap_i\|^2\ ,\tag{6}$$

$$R_{\text{DNN2}}=\underset{\theta2\in\Theta}{\arg\min}\,\|R_{\theta2}(Ap_i)-P_i\|^2\ ,\tag{7}$$

where R_{DNN1} and R_{DNN2} are the mapping functions of DNN1 and DNN2, $\theta1$ and $\theta2$ denote the weights and biases of DNN1 and DNN2, respectively. Once the learning step is completed, R_{DNN1} and R_{DNN2} can then be used to decipher the plaintext from an unknown ciphertext.

Data acquisition

For training the DNNs mentioned above, the training data should be prepared. The objective of DNN1 is to remove the speckle noise from the autocorrelation functions of ciphertexts. In the COA scenario, the intercepted ciphertext might be encrypted with any unknown RPMs. To get the better de-noising performance, the denoising model should sufficiently encompass the statistical variations across as many RPMs as possible. Usually, different realizations of the speckle patterns can be obtained by coherently illuminating the plaintext with different random phases. This requires the use of many different random phase keys to encrypt the plaintext images to achieve the statistical ergodic property of the speckle pattern. Therefore, a virtual DRPE system (not the real one) should be designed to gather the training data. As illustrated in Fig. 2, a set of randomly generated RPMs are placed at the spatial and frequency domains in this virtual DRPE system to encrypt the plaintext images and obtain the corresponding ciphertext images, which can be expressed as

$$\begin{aligned}C_i(x,y)=&\ \text{FT}^{-1}\{\text{FT}\{P_i(x,y)\\ &\cdot M_i(x,y)\}\cdot N_i(u,v)\},i=1,2,...,n\ .\end{aligned}\tag{8}$$

Subsequently, the autocorrelations of ciphertexts $Ac_i(x,y)$ can be calculated by taking an inverse Fourier transform of its power spectrum in accordance with the Wiener–Khinchin theorem, that is

$$Ac_i(x,y)=\left|\text{FT}^{-1}\{|\text{FT}\{C_i(x,y)\}|^2\}\right|,i=1,2,...,n\ .\tag{9}$$

Meanwhile, the autocorrelations of plaintexts $Ap_i(x,y)$

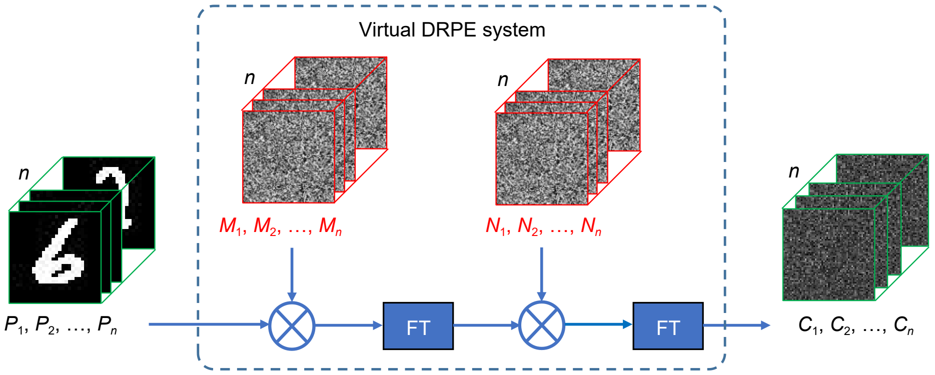

Fig. 2 | Acquisition of the training data by designing a virtual DRPE system. A set of randomly generated RPMs (M_1, M_2,…,M_n) are placed at the spatial domain, and another set of randomly generated RPMs (N_1, N_2,…,N_n) are placed at the frequency domain. The ground truth plaintext images (P_1, P_2,…,P_n) are encrypted one-by-one and the corresponding ciphertext dataset (C_1, C_2,…,C_n) can be obtained.

can also be calculated from the ground-truth plaintext images, that is

$$Ap_i(x,y) = \left|\mathrm{FT}^{-1}\{|\mathrm{FT}\{P_i(x,y)\}|^2\}\right|, \quad i=1,2,\ldots,n\,. \tag{10}$$

In this way, the dataset of the autocorrelations of ciphertexts $Ac_i(x,y)$ and the autocorrelations of plaintexts $Ap_i(x,y)$ are set as the inputs and outputs of DNN1 respectively; the dataset of the autocorrelations of plaintexts $Ap_i(x,y)$ and the ground-truth plaintext images $P_i(x,y)$ are set as the inputs and outputs of DNN2, respectively.

Network model

To perform the task of de-noising, the popular DnCNN model[60] is employed as DNN1 in the proposed COA method. The architecture is illustrated in Fig. 3(a). The ciphertext's autocorrelation image first passes through a standard convolutional layer with filter size 3×3, followed by a rectified linear unit (ReLU), and then passes

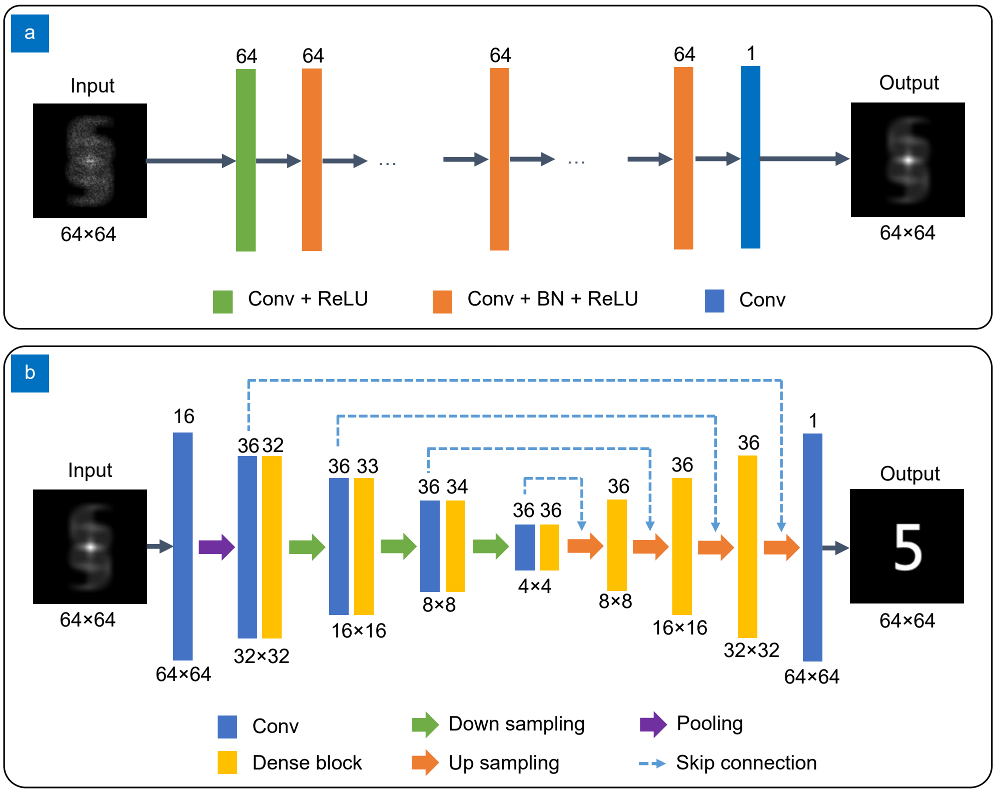

Fig. 3 | Structure of the employed DNNs. (**a**) The architecture of DNN1, which takes the DnCNN structure. (**b**) The architecture of DNN2, which takes the general encoder–decoder U-net structure. The encoder gradually condenses the lateral spatial information into high-level feature maps with growing depths; the decoder reverses the process by recombining the information into feature maps with gradually increased lateral details.

through ten "BN+Conv+ReLU" blocks successively, which consists of an operation of batch normalization (BN), a standard convolutional layer with the size 3×3 and a ReLU. Finally, a standard convolutional layer with filter size 1×1 is employed to output the noise-free autocorrelation image. We have compared DnCNN model with the extensively-utilized U-net one[59] on the task of de-noising (See Fig. S1). Both networks have good performance on the removal of speckle noise, but the training time of DnCNN is much shorter than that of the U-net.

However, for the task of de-correlation, a modified U-net architecture[59] is employed (see Fig. 3(b)), where the encoder path (left side) to extract the feature maps from the input patterns, and the decoder path (right side) to perform pixel-wise regression. The input to the network is the de-noised autocorrelation images with the size 64 pixels ×64 pixels. They first pass through a standard convolutional layer with the filter size 3×3, followed by a 2×2 max pooling operation with the stride 2 for downsampling, and then is successively decimated by four "convolution + dense + downsampling" blocks, where each block consists of a standard convolutional layer with the size 3×3, a dense block, and a 2×2 max pooling layer with the stride 2. A dense block consists of n convolutional layers, where the n^{th} layer receives the feature-maps of all preceding layers, $x_0, x_1, \ldots, x_{n-1}$, as inputs:

$$x_n = H_n([x_0, x_1, \ldots, x_{n-1}]) \ , \qquad (11)$$

where the bracket $[\cdot]$ denotes the concatenation of the feature-maps extracted from layers $0, 1, \ldots, n-1$, and $H_n(\cdot)$ refers to a composite function of three consecutive operations: BN, followed by a ReLU and a standard convolution layer with the size 3×3. After passing through the encoder path, the feature maps of the plaintext then successively pass through 3 "upsampling + dense" blocks, where each block consists of an up-convolutional layer with the size 2×2 and a dense block. Finally, another up convolutional layer is employed to perform pixel-wise regression. In addition, skip connections are also employed to pass high-frequency information learned from previous layers down the network toward the output reconstruction. We have also compared two models above on the task of de-correlation, the results are shown in Fig. S2. It is obvious that U-net has better performance than DnCNN for de-correlation.

Results and discussion

Simulations and analysis

Numerical simulations have been carried out to demonstrate the validity of the proposed learning-based COA approach. In the following numerical simulations, the size of all the images is set as 64 pixels ×64 pixels. For the training process, a total of 10000 images (5000 digits and 5000 letters) from the MNITS handwritten digit dataset[61] and NIST handwritten letter[62] were adopted as the dataset of plaintexts. They were resized from 28 pixels ×28 pixels to 32 pixels ×32 pixels and zero-padded 64 pixels ×64 pixels. Simultaneously, 20000 RPMs randomly distributed in the range of (0,2π] with 256 gray-levels were generated by setting the different random seeds (1–10000 for RPM1, 10001–20000 for RPM2), which were used as the phase keys at the spatial and frequency domains of the DRPE system respectively. Subsequently, 10000 corresponding ciphertexts were obtained by encrypting the 10000 plaintext images with this virtual DRPE system. The autocorrelations of ciphertexts and plaintexts were calculated as the input and output of the DNN1, and the autocorrelations of plaintexts and plaintexts itself acted as the input and output of the DNN2. For both DNNs, the loss functions have been defined by the mean absolute error (MAE):

$$MAE(A, B) = \frac{1}{N} \sum_{i,j} [A(i,j) - B(i,j)] \ , \qquad (12)$$

where $A(i,j)$ and $B(i,j)$ denote the output of the DNNs and the ground truth, respectively, N is the number of the pixels. Alternatively, a loss function that is based on mean square error (MSE) can also be used to obtain similar results. We used the Adam optimizer with a learning rate of 0.0005 to optimize the weights and biases of the neural networks. The program was implemented with Python 3.6 on the platform of TensorFlow. A graphics processing unit (NVIDIA GeForce GTX 1050 Ti) was used to expedite the computation. After 10 epochs and 20 epochs, the loss function MAEs of DNN1 and DNN2 become 0.0037 and 0.048, respectively, which implies that the DNNs have been well trained to achieve a good performance for the training dataset.

With the two trained DNNs at hand, now we can perform the COA test. The numerical simulation results are shown in Fig. 4, where the three columns on the left indicate the digits while another three columns on the right show letters. Figure 4(a) shows the given ciphertexts, which are generated with the testing plaintext images (different from the training dataset) in a testing DRPE system (RPMs generated by setting the random seeds as 20001–21000 for RPM1 and 21001–22000 for RPM2). With the given ciphertexts, we can calculate

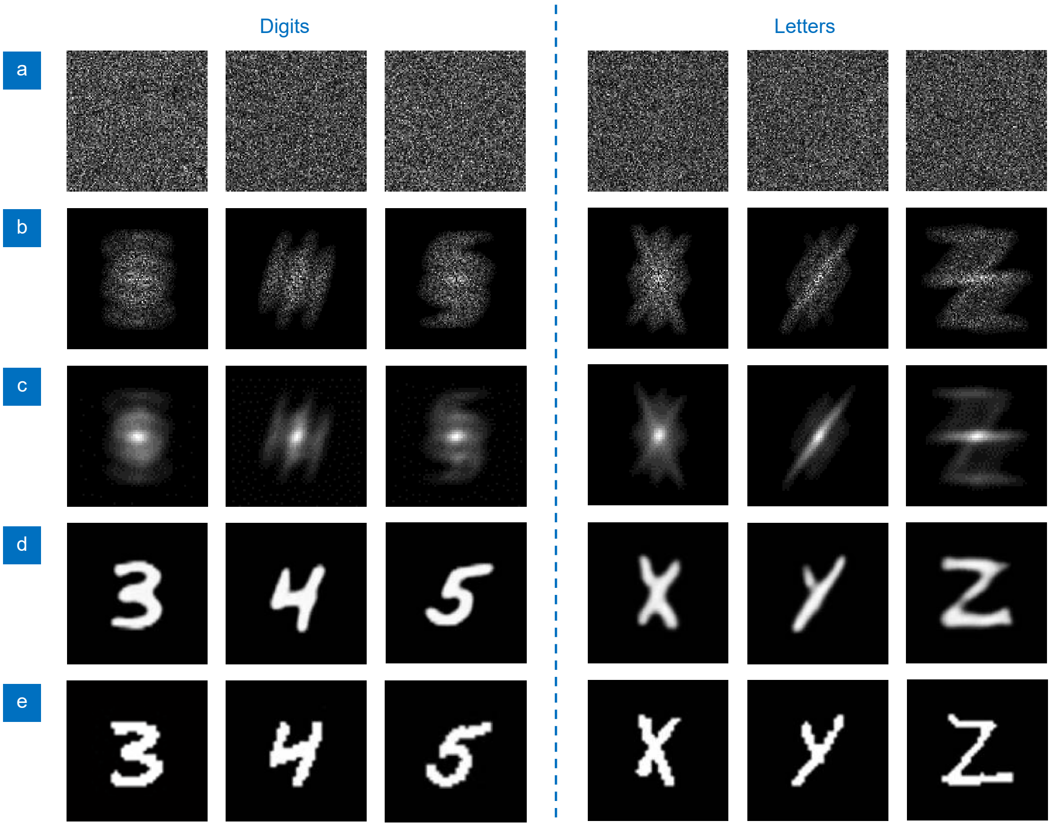

Fig. 4 | Attack results by our proposed COA approach. (**a**) The given ciphertexts. (**b**) The autocorrelations of ciphertexts. (**c**) Outputs of DNN1. (**d**) Outputs of DNN2. (**e**) The ground-truth plaintext images.

their autocorrelation functions, which are presented in Fig. 4(b). We have removed the peaked function $\delta(x,y)$ from the autocorrelation images for displaying. It can be clearly seen that the autocorrelation functions present inerratic edge profiles and is corrupted by random speckle noise. Subsequently, these pre-processed autocorrelation functions are fed into the trained DNN1, and the outputs are presented in Fig. 4(c). Then the outputs from DNN1 are further fed into the trained DNN2 model, the outputs of DNN2 are presented in Fig. 4(d). The ground-truth plaintext images are shown in Fig. 4(e) for comparison. The predicted plaintext images can be visualized and are clearly recognizable despite of the fact that their resolutions are slightly degraded which may be resulted from the convolution operation in the DNN training process.

It should be pointed out that zero-padding of images was applied before the encryption to introduce frequency redundancy. Zero-padding operation actually has been extensively exploited and discussed in signal processing literature[28–33]. According to Nyquist-Shannon theorem, a two-dimensional signal can be uniquely specified by the magnitude of its twice oversampled discrete Fourier transformat[34]. Conventional phase retrieval methods do work well unless the condition of twice oversampling is satisfied. To further validate the proposed method, we try to reconstruct plaintext image from its autocorrelation function without zero-padding operation, which is usually impossible for the conventional iterative phase retrieval algorithm. During the training process, we use the autocorrelation function without zero-padding as the input of the de-correlation DNN model and the test results are presented in Fig. S4. Surprisingly, the images can be recovered with high fidelity from the incomplete autocorrelation pattern, which indicates the de-correlation DNN model still effective even without zero-padding operation.

To quantitatively analyze the reliability of the proposed COA method, we introduce the correlation coefficient (CC) to quantitatively evaluate the quality of the retrieved plaintext images. The CC between image A and image B are defined as follows

$$CC(A,B)=\frac{\sum_{i,j}[A(i,j)-\overline{A}][B(i,j)-\overline{B}]}{\sqrt{\sum_{i,j}[A(i,j)-\overline{A}]^2\times\sum_{i,j}[B(i,j)-\overline{B}]^2}}\ , \tag{13}$$

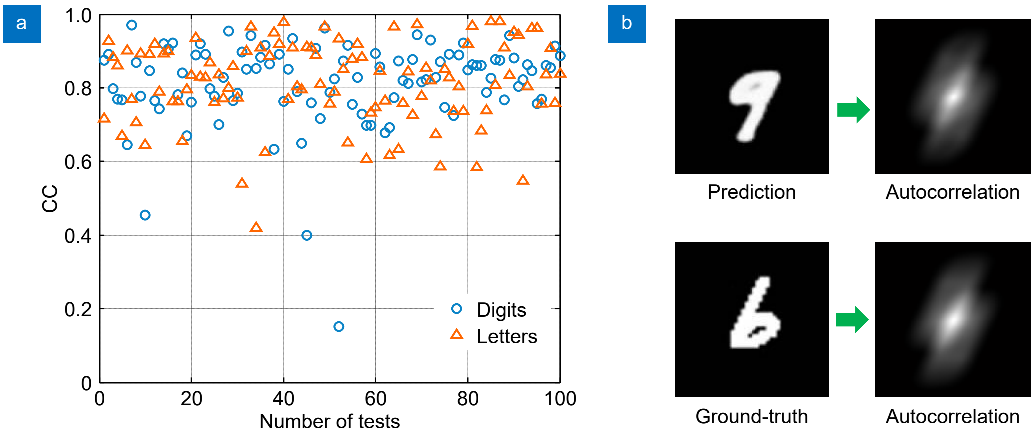

Fig. 5 | Quantitative evaluation of the reliability of the proposed COA method. (**a**) CC values to the number of tests. (**b**) The example of the prediction rotated 180 degrees and the ground-truth, which have the similar autocorrelation.

where $\bar{A}$ and $\bar{B}$ denote average value of $A(i,j)$ and $B(i,j)$. We have calculated the values of the CC of 100 reconstructed plaintext images which are randomly selected out of the total 1000 in the test set, and the results are plotted in Fig. 5(a). The blue circle and orange triangle markers represent the values of the CC related to digits and letters test data, respectively. As expected, most of the CC values are larger than 0.5 and the averaged CC values is 0.816. However, it is also worth noting that there are still a few cases in which the CC values are smaller than 0.5 due to the fact that the retrieved plaintext images rotated 180 degrees, especially for the digits "6" and "9". The reason for that is the autocorrelations of the prediction and the ground-truth are almost the same, as shown in Fig. 5(b). The results consist of the ambiguity regarding the phase-retrieval inverse problem since the recovery of a signal from its Fourier magnitude alone, in general, does not yield a unique solution[34].

Moreover, we have also investigated the robustness of the proposed method against the cropping and the noise. The results are illustrated in Fig. 6. The two images on the left side of Fig. 6(a) respectively present the cropped ciphertexts with cropping ratio 1/16 and 1/4; the two images in the middle of Fig. 6(a) present the ciphertexts

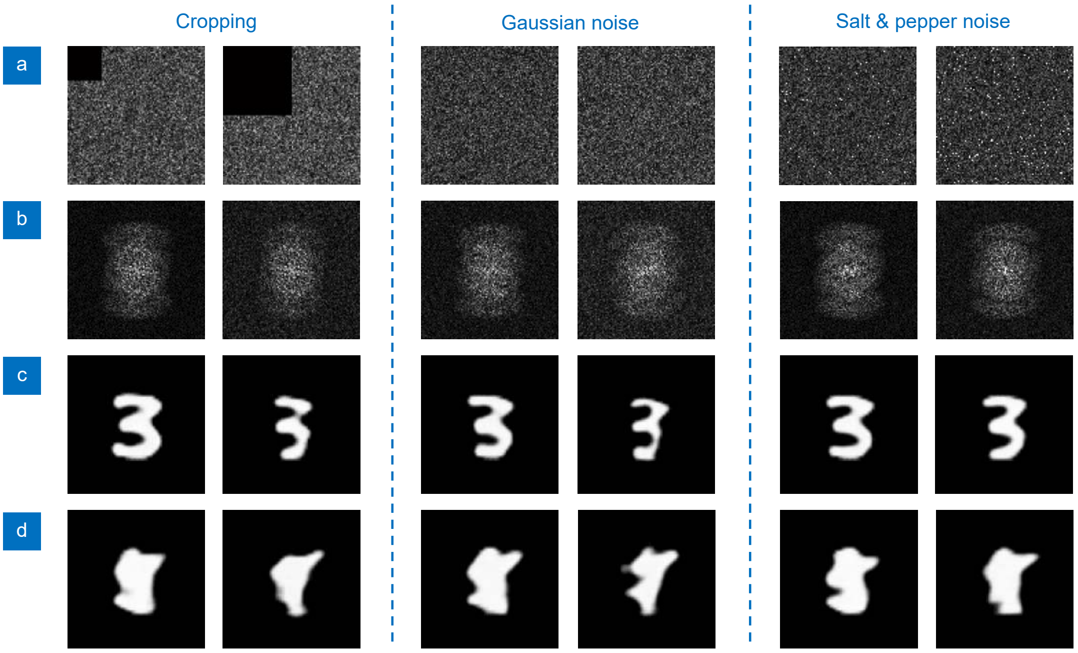

Fig. 6 | Robustness test against cropping and noise. (**a**) Ciphertexts with cropping ratio 1/16 and 1/4, added zero-mean Gaussian noise with 0.01 and 0.02 variance, and added salt & pepper noise with 0.01 and 0.02 distribution density. (**b**) The corresponding autocorrelation distributions. (**c**) The retrieved images by the proposed two-step learning-based COA method. (**d**) The retrieved images by the one-step learning-based method.

Table 1 | CC values between the retrieved plaintexts and the ground-truth

Methods	Cropping		Gaussian noise		Salt & pepper noise	
Two-step method	0.9464	0.7522	0.8958	0.7610	0.9284	0.8038
One-step method	0.4517	0.3351	0.3751	0.2914	0.4116	0.3290

added zero-mean Gaussian noise with 0.01 and 0.02 variance; the two images on the right side of Fig. 6(a) present the ciphertexts added salt & pepper noise with 0.01 and 0.02 distribution density. The calculated autocorrelation functions of the corrupted ciphertexts are displayed in Fig. 6(b). The reconstructed plaintext images by the proposed two-step deep-learning-based COA method are shown in Fig. 6(c). For comparison, the reconstructed images by the one-step "end-to-end" method (from the autocorrelation of ciphertext to the plaintext directly) are shown in Fig. 6(d). Obviously, the images shown in Fig. 6(c) can be visualized and recognized while the images shown in Fig. 6(d) are completely different from the ground-truth. To quantitatively evaluate the robust capability, we have calculated the CC between the retrieved images (Figs. 6(c) and 6(d)) and the ground-truth image (the first image from the left of Fig. 4(e)), the CC values are shown in Table 1. More data on CC values under various levels of cropping and noise were presented in Fig. S3. These results indicate that the proposed method has the better robustness against the cropping and the noise than the one-step method.

Optical experiments

To further experimentally verify the effectiveness and practicability of the proposed learning-based COA approach, we designed and set up an experiment configuration that is schematically shown in Fig. 7. A continuous-wave laser (MW-SL-532/50mW) served as the illumination source. A spatial filter and a collimating lens were placed behind the laser. A spatial light modulator (SLM) (Holoeye LC2002, transmission) was placed at the input plane to display the plaintext images. Two orthogonally oriented polarizers were placed before and after the SLM to ensure that the SLM worked in amplitude mode. A thin diffuser served as the RPM was placed next to the SLM. A high dynamic range CMOS camera (PCO edge 4.2, 2160 pixels ×2160 pixels with a pixel size of 6.5 μm × 6.5 μm, dynamic range of 16 bits) was placed on the back focal plane of the Fourier lens (f = 150 mm) to capture the power spectrum. Considering the effect of the RPM2 could be removed by the autocorrelation operation, we do not set the second RPM2 at frequency plane in the following experiments.

In the training process, 1000 images (28 pixels ×28 pixels) from the Quickdraw dataset[63] were selected as the plaintext samples. After scaling to the size of 100 pixels ×100 pixels, they were zero-padded to 1024 pixels × 768 pixels and loaded onto the SLM in sequence. A commercial ground glass as the training diffuser (radius r = 50 mm) rotating constantly at 1 circle per minute, which can provide about 600 different RPMs. Subsequently, 1000 corresponding power spectrum images were collected by the camera. For the test process, another 1000 images excluded from the training dataset were uploaded onto the SLM as testing plaintexts. The ground-truth plaintext images "apple" and "banana" as the representative examples of the testing images. We used three different types of diffuser (DG600, DG220 and DG120, Thorlabs) to capture the power spectrum images as testing ciphertexts. Figure 8 shows some of experimental test results. The measured power spectra images are presented

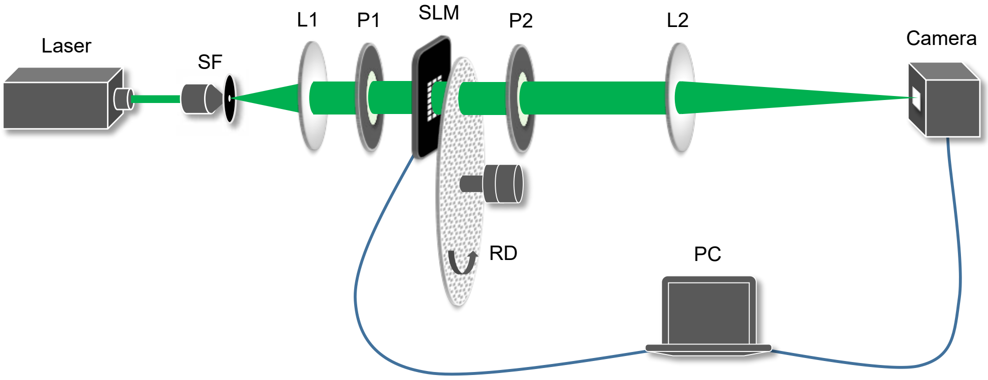

Fig. 7 | Experimental setup. SF: spatial filter, L1: collimating lens, L2: Fourier lens, P1 and P2: polarizers, RD: rotating diffuser, SLM: spatial light modulator.

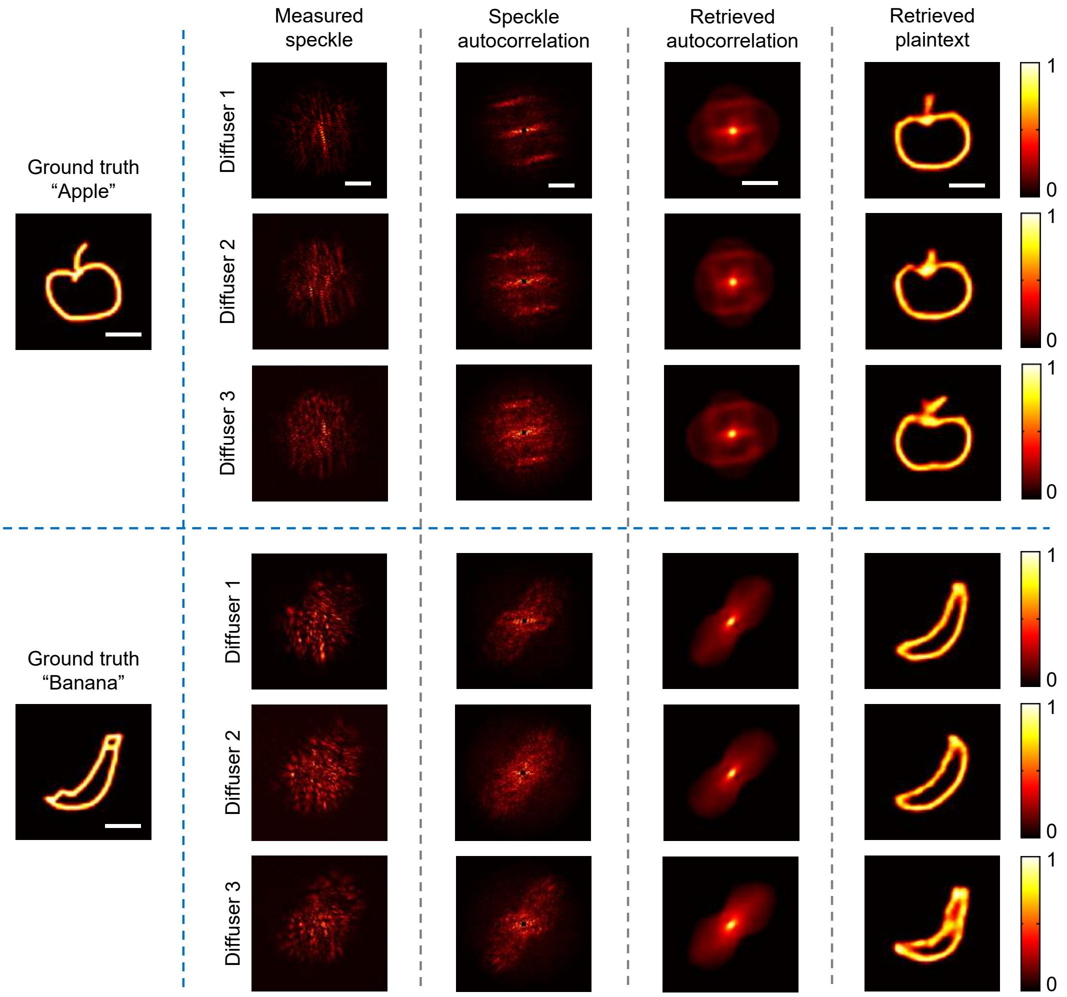

Fig. 8 | Experimental results. First column from left: the ground truth plaintext images "Apple" and "Banana". Second column: the raw power spectrum images of the different types of diffusers. Third columns: the corresponding autocorrelation functions. Fourth and fifth columns: the recovered autocorrelation from DNN1 and the retrieved plaintext images from DNN2. Scale bar: 100 pixels in pictures of the second and third columns from left; 20 pixels in pictures of the first, fourth and fifth columns from left.

in the second column from left. These three pictures appear to be obviously different although they are from the same plaintext. However, the autocorrelations of these images reveal the similar patterns. The retrieved autocorrelations from DNN1 and the retrieved plaintext images are shown in the fourth and fifth columns of Fig. 8. It is suggested that the proposed method consistently makes high-quality plaintexts retrieval from the ciphertexts of different DRPE systems. We have calculated the CC values of 1000 test data between the reconstructed plaintext images and the ground-truth images, and the average CC values of three diffusers are 0.7635, 0.6969 and 0.6214, respectively.

Conclusions

In summary, we have developed a two-step deep learning strategy and demonstrated numerically and experimentally that it is capable of achieving COA on the classical DRPE system. By incorporating the deep learning method with the speckle correlation technique, the proposed learning-based COA scheme employs two DNNs to respectively learn the removal of speckle noise in the autocorrelation domain and the de-correlation operation for deciphering plaintext images. Compared with existing learning-based attack methods, the proposed method has a unique character that the mapping relationships of autocorrelation features are trained, instead of the random phase keys of DRPE system so that our approach allows to retrieve the plaintext from the only ciphertext without any other resources. Furthermore, the proposed COA method can be very efficient because the plaintext can be retrieved from the intercepted ciphertext in real-time with use of the trained DNNs. One of limitations of the proposed method is that the capacity of the generalization of de-correlation DNN model is limited, and this COA approach works well only when the test images are similar to those in the training dataset. Therefore, it should be better if the training dataset includes more types of plaintext images since the training process of two DNNs can be done before the real COA process.

References

1. Javidi B, Carnicer A, Yamaguchi M, Nomura T, Pérez-Cabré E et al. Roadmap on optical security. *J Opt* **18**, 083001 (2016).
2. Carnicer A, Javidi B. Optical security and authentication using nanoscale and thin-film structures. *Adv Opt Photonics* **9**, 218 (2017).
3. Refregier P, Javidi B. Optical image encryption based on input plane and Fourier plane random encoding. *Opt Lett* **20**, 767–769 (1995).
4. Unnikrishnan G, Joseph J, Singh K. Optical encryption by double-random phase encoding in the fractional Fourier domain. *Opt Lett* **25**, 887–889 (2000).
5. Zhu BH, Liu ST, Ran QW. Optical image encryption based on multifractional Fourier transforms. *Opt Lett* **25**, 1159–1161 (2000).
6. Situ GH, Zhang JJ. Double random-phase encoding in the Fresnel domain. *Opt Lett* **29**, 1584–1586 (2004).
7. Mehra I, Nishchal NK. Image fusion using wavelet transform and its application to asymmetric cryptosystem and hiding. *Opt Express* **22**, 5474–5482 (2014).
8. Javidi B, Nomura T. Securing information by use of digital holography. *Opt Lett* **25**, 28–30 (2000).
9. Kong DZ, Cao LC, Shen XJ, Zhang H, Jin GF. Image encryption based on interleaved computer-generated holograms. *IEEE Trans Ind Inform* **14**, 673–678 (2018).
10. Nomura T, Javidi B. Optical encryption using a joint transform correlator architecture. *Opt Eng* **39**, 2031–2035 (2000).
11. Zhang Y, Wang B. Optical image encryption based on interference. *Opt Lett* **33**, 2443–2445 (2008).
12. Chen W, Chen XD, Sheppard CJR. Optical image encryption based on diffractive imaging. *Opt Lett* **35**, 3817–3819 (2010).
13. Clemente P, Durán V, Torres-Company V, Tajahuerce E, Lancis J. Optical encryption based on computational ghost imaging. *Opt Lett* **35**, 2391–2393 (2010).
14. Shi YS, Li T, Wang YL, Gao QK, Zhang SG et al. Optical image encryption via ptychography. *Opt Lett* **38**, 1425–1427 (2013).
15. Schneier B. *Applied Cryptography*: *Protocols*, *Algorithms*, *and Source Code in C* 2nd ed (Wiley, New York, 1996).
16. Cheng XC, Cai LZ, Wang YR, Meng XF, Zhang H et al. Security enhancement of double-random phase encryption by amplitude modulation. *Opt Lett* **33**, 1575–1577 (2008).
17. Liao MH, He WQ, Lu DJ, Wu JC, Peng X. Security enhancement of the phase-shifting interferometry-based cryptosystem by independent random phase modulation in each exposure. *Opt Lasers Eng* **89**, 34–39 (2017).
18. Sahoo SK, Tang DL, Dang C. Enhancing security of incoherent optical cryptosystem by a simple position-multiplexing technique and ultra-broadband illumination. *Sci Rep* **7**, 17895 (2017).
19. Peng X, Wei HZ, Zhang P. Chosen-plaintext attack on lensless double-random phase encoding in the Fresnel domain. *Opt Lett* **31**, 3261–3263 (2006).
20. Liao MH, Lu DJ, He WQ, Peng X. Optical cryptanalysis method using wavefront shaping. *IEEE Photonics J* **9**, 2200513 (2017).
21. Peng X, Zhang P, Wei HZ, Yu B. Known-plaintext attack on optical encryption based on double random phase keys. *Opt Lett* **31**, 1044–1046 (2006).
22. Gopinathan U, Monaghan DS, Naughton TJ, Sheridan JT. A known-plaintext heuristic attack on the Fourier plane encryption algorithm. *Opt Express* **14**, 3181–3186 (2006).
23. Peng X, Tang HQ, Tian JD. Ciphertext-only attack on double random phase encoding optical encryption system. *Acta Phys Sin* **56**, 2629–2636 (2007).
24. Zhang CG, Liao MH, He WQ, Peng X. Ciphertext-only attack on a joint transform correlator encryption system. *Opt Express* **21**, 28523–28530 (2013).
25. Liu XL, Wu JC, He WQ, Liao MH, Zhang CG et al. Vulnerability to ciphertext-only attack of optical encryption scheme based on double random phase encoding. *Opt Express* **23**, 18955–18968 (2015).
26. Fienup JR. Reconstruction of an object from the modulus of its Fourier transform. *Opt Lett* **3**, 27–29 (1978).
27. Fienup JR. Phase retrieval algorithms: a comparison. *Appl Opt* **21**, 2758–2769 (1982).
28. Hayes M, Lim J, Oppenheim A. Signal reconstruction from phase or magnitude. *IEEE Trans Acoust Speech Signal Process* **28**, 672–680 (1980).
29. Michael G, Porat M. On signal reconstruction from Fourier magnitude. In *Proceedings of the 8th IEEE International Conference on Electronics*, *Circuits and Systems* 1403–1406 (IEEE, 2001). https://doi.org/10.1109/ICECS.2001.957477.
30. Sarang R, Motlagh MRJ, Eslami P. Reconstruction of image using just magnitude information of Fourier transform; is phase information really more important? In *Proceedings of 2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce* 56–56 (IEEE, 2006). http://doi.org/10.1109/CIMCA.2006.172.
31. Isernia T, Pascazio V, Pierri R, Schirinzi G. Image reconstruction from Fourier transform magnitude with applications to synthetic aperture radar imaging. *J Opt Soc Am A* **13**, 922–934 (1996).
32. Gerchberg RW, Saxton WO. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972).
33. Griffin D W, Lim J S. Signal estimation from modified short-time Fourier transform. *IEEE Trans Acoust Speech Signal Process* **32**, 236–243 (1984).
34. Shechtman Y, Eldar YC, Cohen O, Chapman HN, Miao JW et al. Phase retrieval with application to optical imaging: a contemporary overview. *IEEE Signal Process Mag* **32**, 87–109 (2015).
35. Liao MH, He WQ, Lu DJ, Peng X. Ciphertext-only attack on optical cryptosystem with spatially incoherent illumination: from the view of imaging through scattering medium. *Sci Rep* **7**, 41789 (2017).

36. Liao MH, Lu DJ, He WQ, Peng X. Speckle-correlation-based ciphertext-only attack on the double random phase encoding scheme. *Proc SPIE* **10250**, 102502i (2017).
37. Li GW, Yang WQ, Li DY, Situ GH. Cyphertext-only attack on the double random-phase encryption: experimental demonstration. *Opt Express* **25**, 8690–8697 (2017).
38. Barbastathis G, Ozcan A, Situ GH. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
39. Chen LW, Yin YM, Li Y, Hong MH. Multifunctional inverse sensing by spatial distribution characterization of scattering photons. *Opto-Electron Adv* **2**, 190019 (2019).
40. Saetchnikov AV, Tcherniavskaia EA, Saetchnikov VA, Ostendorf A. Deep-learning powered whispering gallery mode sensor based on multiplexed imaging at fixed frequency. *Opto-Electron Adv* **3**, 200048 (2020).
41. Kamilov US, Papadopoulos IN, Shoreh MH, Goy A, Vonesch C et al. Learning approach to optical tomography. *Optica* **2**, 517–522 (2015).
42. Lyu M, Wang W, Wang H, Wang HC, Li GW et al. Deep-learning-based ghost imaging. *Sci Rep* **7**, 17865 (2017).
43. Wang F, Wang H, Wang HC, Li GW, Situ GH. Learning from simulation: an end-to-end deep-learning approach for computational ghost imaging. *Opt Express* **27**, 25560–25572 (2019).
44. Zuo HR, Xu ZY, Zhang JL, Jia G. Visual tracking based on transfer learning of deep salience information. *Opto-Electron Adv* **3**, 190018 (2020).
45. Rivenson Y, Zhang YB, Günaydın H, Teng D, Ozcan A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci Appl* **7**, 17141 (2018).
46. Wang H, Lyu M, Situ GH. eHoloNet: a learning-based end-to-end approach for in-line digital holographic reconstruction. *Opt Express* **26**, 22603–22614 (2018).
47. Sinha A, Lee J, Li S, Barbastathis G. Lensless computational imaging through deep learning. *Optica* **4**, 1117–1125 (2017).
48. Cherukara MJ, Nashed YSG, Harder RJ. Real-time coherent diffraction inversion using deep generative networks. *Sci Rep* **8**, 16520 (2018).
49. Wang F, Bian YM, Wang HC, Lyu M, Pedrini G et al. Phase imaging with an untrained neural network. *Light Sci Appl* **9**, 77 (2020).
50. Li S, Deng M, Lee J, Sinha A, Barbastathis G. Imaging through glass diffusers using densely connected convolutional networks. *Optica* **5**, 803–813 (2018).
51. Li YZ, Xue YJ, Tian L. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica* **5**, 1181–1190 (2018).
52. Metzler CA, Heide F, Rangarajan P, Balaji MM, Viswanath A et al. Deep-inverse correlography: towards real-time high-resolution non-line-of-sight imaging. *Optica* **7**, 63–71 (2020).
53. Lyu M, Wang H, Li GW, Zheng SS, Situ GH. Learning-based lensless imaging through optically thick scattering media. *Adv Photonics* **1**, 036002 (2019).
54. Hai H, Pan SX, Liao MH, Lu DJ, He WQ et al. Cryptanalysis of random-phase-encoding-based optical cryptosystem via deep learning. *Opt Express* **27**, 21204–21213 (2019).
55. Zhou LN, Xiao Y, Chen W. Machine-learning attacks on interference-based optical encryption: experimental demonstration. *Opt Express* **27**, 26143–26154 (2019).
56. Zhou LN, Xiao Y, Chen W. Vulnerability to machine learning attacks of optical encryption based on diffractive imaging. *Opt Lasers Eng* **125**, 105858 (2020).
57. Qin Y, Wan YH, Gong Q. Learning-based chosen-plaintext attack on diffractive-imaging-based encryption scheme. *Opt Lasers Eng* **127**, 105979 (2020).
58. Zhou LN, Xiao Y, Chen W. Learning-based attacks for detecting the vulnerability of computer-generated hologram based optical encryption. *Opt Express* **28**, 2499–2510 (2020).
59. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention* 234 –241 (Springer, 2015). http://doi.org/10.1007/978-3-319-24574-4_28.
60. Zhang K, Zuo WM, Chen YJ, Meng DY, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Trans Image Process* **26**, 3142–3155 (2017).
61. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* **86**, 2278–2324 (1998).
62. Garris MD, Blue JL, Gerald TC, Grother PJ, Wilson CL. NIST Form-Based Handprint Recognition System (US Department of Commerce, Technology Administration, National Institute of Standards and Technology: Gaithersburg, MD, USA, 1997).
63. Ha D, Eck D. A neural representation of sketch drawings. arXiv: 1704.03477 (2017).

Acknowledgements

We are grateful for financial supports from the National Natural Science Foundation of China (NSFC) (62061136005, 61705141, 61805152, 61875129, 61701321); Sino-German Research Collaboration Group (GZ 1391) and the Mobility program (M-0044) sponsored by the Sino-German Center; Chinese Academy of Sciences (QYZDB-SSW-JSC002); Science and Technology Innovation Commission of Shenzhen (JCYJ20170817095047279).

Author contributions

M. Liao and S. Zheng contributed equally to this work. M. Liao, S. Zheng and G. Situ conceived the design, and discussed the experimental implementation. M. Liao, S. Zheng and S. Pan preformed the simulations and experiments. D. Lu, W. He, G. Situ and X. Peng discussed the results. X. Peng and G. Situ conceived the idea and supervised the project. All authors contributed to the manuscript.

Competing interests

The authors declare no competing financial interests.

Supplementary information

Supplementary information for this paper is available at https://doi.org/10.29026/oea.2021.200016