Optics EXPRESS

Ghost translation: an end-to-end ghost imaging approach based on the transformer network

WENHAN REN,[1] XIAOYU NIE,[2] TAO PENG,[1,*] AND MARLAN O. SCULLY[1,3,4,5]

[1] *Institute for Quantum Science and Engineering, Texas A&M University, College Station, Texas, 77843, USA*
[2] *School of Physics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*
[3] *Baylor University, Waco, 76706, USA*
[4] *Princeton University, Princeton, New Jersey 08544, USA*
[5] *scully@tamu.edu*
[*] *taopeng@tamu.edu*

Abstract: Artificial intelligence has recently been widely used in computational imaging. The deep neural network (DNN) improves the signal-to-noise ratio of the retrieved images, whose quality is otherwise corrupted due to the low sampling ratio or noisy environments. This work proposes a new computational imaging scheme based on the sequence transduction mechanism with the transformer network. The simulation database assists the network in achieving signal translation ability. The experimental single-pixel detector's signal will be 'translated' into a 2D image in an end-to-end manner. High-quality images with no background noise can be retrieved at a sampling ratio as low as 2%. The illumination patterns can be either well-designed speckle patterns for sub-Nyquist imaging or random speckle patterns. Moreover, our method is robust to noise interference. This translation mechanism opens a new direction for DNN-assisted ghost imaging and can be used in various computational imaging scenarios.

1. Introduction

Unlike conventional imaging, computational imaging (CI) utilizes both of the optical design and post-detection signal processing to create novel imaging systems. CI frequently outperforms the conventional approach, in terms of imaging resolution, speed, and the signal-to-noise ratio. It has been widely implemented in biophysics [1], tomography [2,3], chemical microscopy [4,5], and non-invasive imaging in complex media [6]. One branch of CI is combining single-pixel detection technique, typically the computational ghost imaging (CGI) [7,8], with reconstruction algorithms to indirectly retrieve the object. The ghost image of an object is not directly obtained through a camera, but it is reconstructed through the spatial correlation of two light beams [9–11]. In the computational fashion, specific speckle patterns are used for the spatial illumination [12,13]. It can also be done on the detected signals in time domain [14,15], and compressive sensing (CS) on the sparsity properties [16,17]. These techniques provide with advantages, *etc.*, high sampling efficiency [18], high-resolution [19,20], broadband accessibility [21,22], and non-optical availability [23,24].

With the prosperity and development of artificial intelligence in recent years, deep learning (DL) has been widely used for CI to achieve extremely low sampling ratios [25–27], or superresoltuion imaging [28,29]. Deep neural networks (DNN) are used in the DL framework to improve the retrieved image quality after training using experimental [30] or simulated [31] data. Despite a variety of realizations, most DL imaging methods are based on a convolutional neural network(CNN). CNN is well-known for extracting the underlying features and visual structure [32]. The application of CNN to CI relies on using the kernel to extract local information over spatial neighborhoods of the imaging object. When the network is well-trained, it can extract the

#478695 https://doi.org/10.1364/OE.478695
Journal © 2022 Received 21 Oct 2022; revised 2 Dec 2022; accepted 3 Dec 2022; published 15 Dec 2022

local features of the otherwise blurred image and give a clearer image with a higher signal-to-noise ratio. However, the underneath features of an object are not local, as suggested by the Fourier spectrum. Therefore, some essential features of the objects are ignored or less emphasized in the standard DL imaging schemes, especially in the sub-Nyquist condition, limiting their application to a broader range.

In this work, we propose and demonstrate a new CI algorithm based on the translation mechanism through the transformer neural network, namely, ghost translation (GT). Unlike the recurrent neural network, in which the data is passed sequentially, The transformer framework processes the entire input all at once [33]. It adopts the self-attention mechanism, deferentially weighting the significance of each part of the input data. The transformer has had great success in natural language processing [34]. Most recently, it has also been demonstrated with excellent performance in the image recognition area [35]. In the spirit of the natural language process, the two variables in the deep learning ghost imaging processes are the bucket signal and the image to be retrieved. The information of the image, especially the relationship between different pixels, is embedded into the bucket signal through the illumination of the speckle patterns. On the other hand, the speckle patterns remain unchanged throughout the process and thus can be considered invariant. To explore and emphasize the relationship between the two variables, we wish to develop an end-to-end approach that can reconstruct the object image directly from the bucket signal without employing illumination patterns. This work implements the translation of light intensity sequences into 2D images. Its unique attention mechanism enables each light intensity to pay attention to the information contained in other light intensities to enrich the connotation. Both simulations and experiments are demonstrated using two types of speckle patterns with different statistical properties. Results from the conventional CGI and CS methods using the same data sets are also presented for comparison. We also analyzed the resulting image quality with three typical evaluating indicators, and all the results suggest that GT outperforms the other methods to a great extent. Indeed, high-quality images can be retired through GT at a sampling ratio as low as 2%. The framework also offers the ability to work in noisy environments.

2. Method

2.1. Theoretical approach

In a typical CI process, K designed patterns P_i (i = 1, 2, 3, $\cdots$, K) are illuminated onto the object $O(x, y)$ in sequence, the transmitted or reflected light intensities I_i are then collected by a bucket detector:

$$I_i = \int P_i(x, y)O(x, y)dxdy. \tag{1}$$

With the bucket signals, the GI can be reconstructed by the correlation function [36]

$$\begin{aligned} O'(x, y) &= \sum_{i=1}^{K} \langle \Delta I_i \Delta P_i(x, y)\rangle \\ &\equiv \sum_{i=1}^{K} \langle I_i - \langle I_i\rangle\rangle\langle P_i(x, y) - \langle P_i(x, y)\rangle\rangle, \end{aligned} \tag{2}$$

where O'(x,y) is the reconstructed object and $\langle\cdot\rangle$ denotes the ensemble average.

As shown in the above equation, although the bucket signal which contains the object information is only 1D, the 2D object can only be retrieved using a sequence of 2D speckle patterns. With CI techniques, a 2D object can be reconstructed from 1D data. The CS reconstruction algorithms search for the most sparse image in the compressible basis, it requires solving a convex optimization program, seeking for the image $\widetilde{O}_{cs}(x, y)$ which minimizes the L_1

norm in the sparse basis [17]

$$arg\ \min||\Psi\{\widetilde{O}(x,y)\}||_{L_1}, \tag{3}$$

subject to

$$\int P_i(x,y)\widetilde{O}(x,y)dxdy = S_i, \quad \forall_i = 1,2,\ldots,M. \tag{4}$$

The problem of finding the image with the minimum L_1 norm can be solved efficiently [37]. Deep learning has also been leveraged to solve the image CS reconstruction process [38,39]. It is worth noting that in a recent work, an end-to-end DLCGI scheme was demonstrated to retrieve the 2D images from only 1D bucket signals [25]. The process can be conceptually understood in two steps: the first step is to recover the illumination patterns $P_i(x,y)$, and the second step is to use the recovered patterns and the bucket signals for CGI processing, the results are then optimized by the DL network [40].

In our proposed scheme, we also use only the 1D bucket signals to reconstruct the 2D image. The 1D bucket is used as the input data, and the 2D image is the output. As shown in Fig. 1, the GT process can be represented as

$$O'(x,y) = O\{I_i\}, \tag{5}$$

where O represents the "translation" process of the trained network. The network in this training process can be written as

$$O' = L\{O(x,y), O\{I_i\}\}, \tag{6}$$

where The cross-entropy loss $L(\ ,\ )$ is calculated based on the cross-entropy between the ground truth object $O(x,y)$ and the "translated" image $O\{I_i\}$. The process of minimizing the cross-entropy loss is used to train the network and get the optimized solution [41].

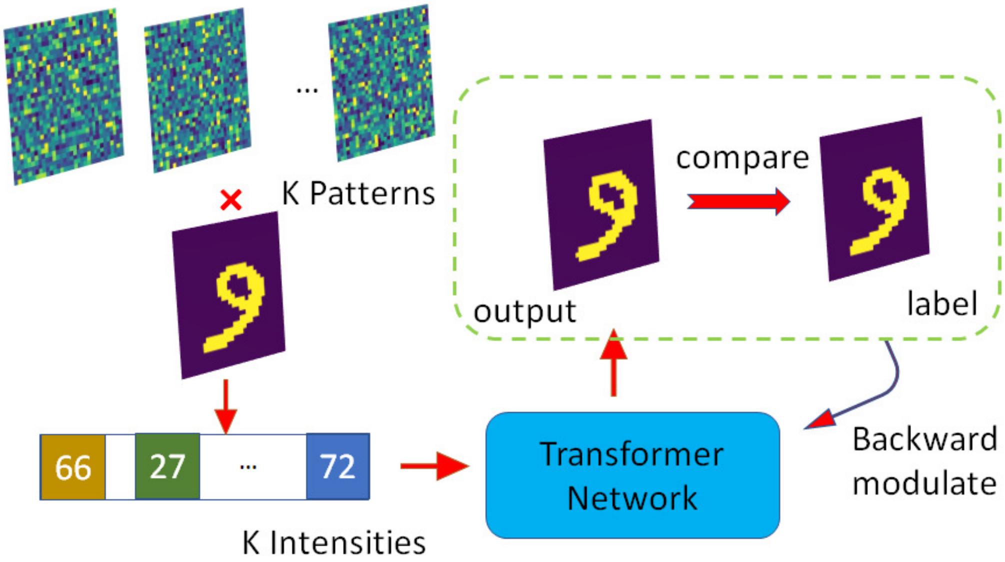

Fig. 1. Schematic of the ghost translation process. Bucket signals are applied to the ghost translation network and being "translated" into a 2D image.

2.2. Transformer network

The transformer is a network based on the Seq2Seq model [42,43]. As its name implies, Seq2Seq converts a sequence of tokens into another sequence of tokens. In general, the Seq2Seq model takes advantage of the Long-Short-Term-Memory(LSTM) model [44], which can be used to give meaning to the sequence, make tokens in the sentence to understand the positional relationship between them, and weaken or strengthen their significance. Seq2Seq is made up of encoders and decoders. The encoders take input tokens and map them to high-dimensional spaces to create richer meanings, like a dictionary for words. The decoders read high-dimensional information generated by encoders and output it in various forms, such as words, characters, or expressions. The transformer is used most frequently in natural language processing because of these characteristics. Following this logic, our work will also treat the light intensity sequence and the two-dimensional image as two different languages. Through the transformer, we can customize the exclusive dictionary for these two languages, so any scene in the intensity sequence can be translated into a 2D image. Our Transformer model, as shown in Fig. 2, uses 6 encoder layers, and 6 decoder layers. We first encode the input sequence to obtain rich embedding for each bucket signal. Each bucket signal value is modeled by a vector of size 512. The first layer of the encoder is the self-attention layer, which is an essential part of the encoder. It can detect the correlation between different light intensity vectors, no matter how far apart. The multi-head mechanism with eight attention heads is also used. These attention computations are then combined to produce the final attention score. This allows our model to discern relationships and nuances between different light intensities. Each decoder layer generates an output sequence by taking all the encoded information. The decoder and encoder layers have residual connections and layer normalization steps, along with feed-forward neural networks for additional processing. In this work, a total of $N = 32 \times 32$ pixel images are used, and each pixel is labeled such that the pixel number in the upper left corner is 1, and it is accumulated in turn, until the grid number in

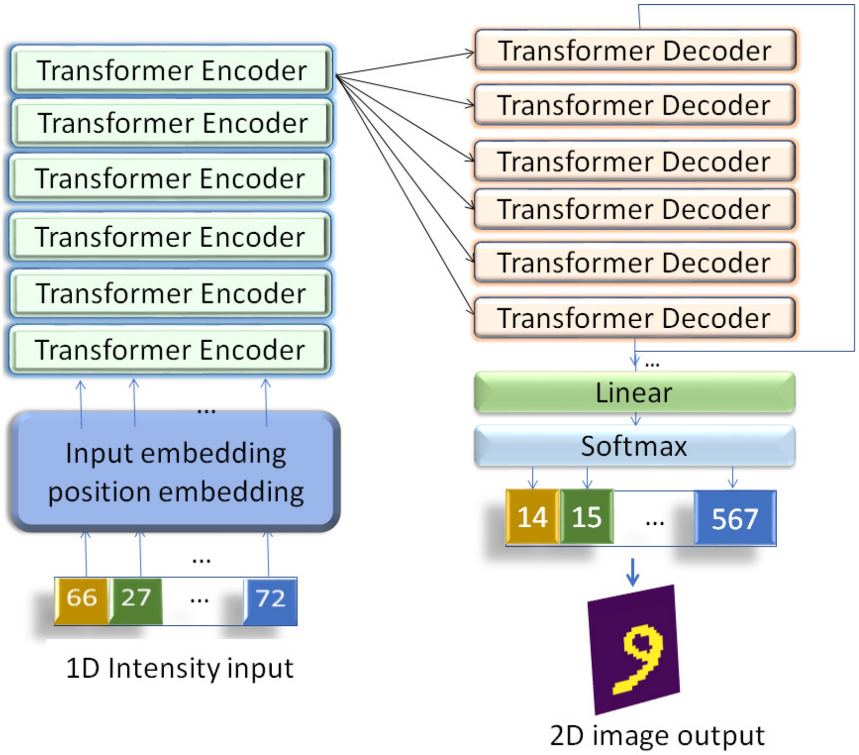

Fig. 2. The transformer network architecture for ghost translation.

the lower right corner is 32×32. Finally, all the grids that light up the numbers are left, resulting in a vector of outputs.

The proposed transformer network requires a training process based on a pre-prepared dataset, and after simulated training, it can directly reconstruct images from light intensity sequences. Here two training sets, i.e., the MNIST handwritten digits database and the Quick Draw smiley face dataset are used. All images are resized to 32×32. During training, we set the batch to 32, train 32 images at a time, and set the maximum epoch to 100. We use the Adam optimizer with a learning rate of 0.001 and use learning rate warm up. We note here that, in the training process, the size of the input vector is fixed, and its size corresponds to the pixel multiples the sampling ratio, so the size of the output is not fixed. Training requires the output size of each batch to be the same, so we will first loop to get the maximum output bucket signal and complete the other outputs to this length. Use padding to 0 for completion, it will be ignored during training [25].

3. Experimental Results

3.1. Experimental setup

Our experimental setup is shown in Fig. 3. As a typical CGI setup, the light source is a He-Ne CW laser with a wavelength of 633 *nm*. The laser is collimated and expanded by the lens system L1 and L2. The beam is then illuminated on the DMD (DLP4100), where the K speckle patterns are displayed sequentially. Each of the K grayscale patterns are first decomposed to 8 monochrome (1-bit, two pixel values) patterns to be compatible with the input format of the DMD [15]. In our experiment, the patterns are filled of 32×32 independent pixels(each one counts 4×4 DMD pixels). The object contains a total of 1024 pixels. The speckle patterns are then projected onto the object plane through a 4f system (L3 and L4). The transmitted light after the object is collected by lens L5 and detected by the single-pixel photodetector (Thorlabs PDA100A2). The single-pixel intensity signal can be used together with the patterns to generate CGI results. It can also be used for the CS algorithm, or for image retrieval though the GT network.

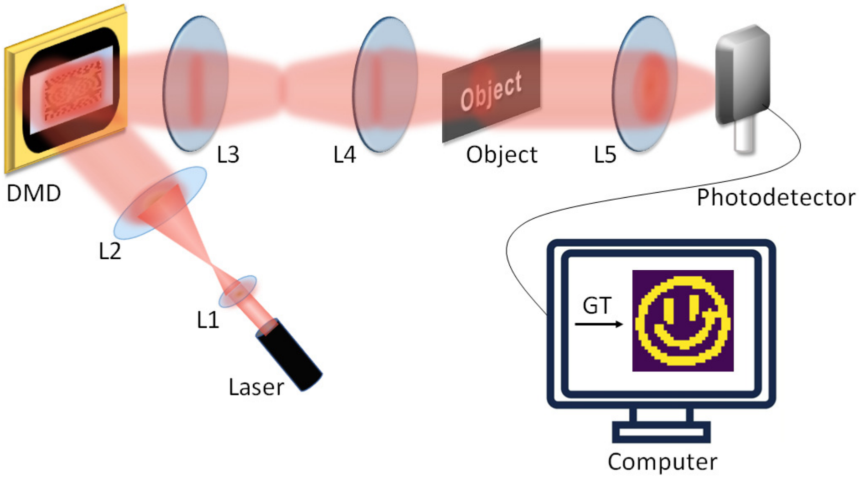

Fig. 3. Schematic setup. L1 and L2 consist of the laser beam expander. The expanded laser beam is modulated by the DMD with a sequence of displayed speckle patterns. The speckle pattern are then mapped onto the object plane through L3 and L4. The transmitted light is collected by a bucket detector which consists of a collection lens (L5) and a single-pixel photodetector. The intensity sequence is sent to the computer for the GT process.

3.2. Results

We used standard Rayleigh speckles with two different sampling ratio of SR=5% (52 speckle patterns) and SR=15% (154 speckle patterns) for the measurements. Ten objects of digits 0-9 from the MNIST handwritten digits database. We note here that these objects are chosen from the testing dataset, which is outside the training dataset. The results are shown in Fig. 4. In Fig. 4(b), we compare the traditional CGI simulation results, the CS simulation and experimental results, GT simulation the experimental GT results at SR=5%. The simulation was done with no noise introduced. The experimental data was taken at a noise level of ∼ 9.6% with a background light source illuminating the detector on purpose. Figure 4(c) presents similar results as of Fig. 4(b) but with SR=15%. For transitional CGI, it is only at 15% sampling ratio when vague and noisy images are retrieve, the images are completely corrupted at the 5% sampling ratio. For the CS algorithm, due to the small pixel sizes (32 × 32), therefore few patterns used, only simulation results at SR of 15% can vague and noisy images be retrieved. When the noise is presented in the experimental data, no images can be retrieved at all. On the other hand, the GT can reconstruct high quality image even at the sampling ratio of 5%, regardless the existence of noise. We notice that there is always no background in the retrieved images, which is another important feature of the GT scheme. To quantitatively justify the quality of reconstructed images, we compare three evaluating indicators of image quality, *i.e.*, the mean square error (MSE), signal-to-noise ratio (SNR), and the structural similarity index measure (SSIM):

$$\begin{aligned}
\mathrm{MSE} &= \frac{1}{N_{\mathrm{pixel}}}\sum_{n=1}^{N_{\mathrm{pixel}}}(O'_i - O_i)^2,\\
\mathrm{SNR} &= 10\log\frac{\sum_n O'_i}{\sum_n |O'_i - O_i|},\\
\mathrm{SSIM(O,O')} &= \frac{(2\bar{O}\bar{O}' + c_1)(2\sigma_{OO'} + c_2)}{(\bar{O}^2 + \bar{O}'^2 + c_1)(\sigma_O^2 + \sigma_{O'}^2 + c_2)}.
\end{aligned} \tag{7}$$

Here $\bar{O}_n$ ($\bar{O}'_n$) is the intensity of the n-th pixel in the object (retrieved image), $\bar{O}$ ($\bar{O}'$) is the mean intensity of the pixels in the corresponding image, σ_O^2 ($\sigma_{O'}^2$) is the variance of the pixel intensity in the corresponding image, $\sigma_{OO'}$ is the covariance of the pixel intensity in both images, c_1 and c_2 are regularization parameters. The MSE and SNR are standard parameters which characterize the image quality in terms of the signal and noise level, as compared to the original object, and SSIM describes changes in structural similarity between the objects, which is more sensitive to human perception [45]. These values are calculated with the ten digits from Fig. 4. As shown in Fig. 5, the image quality of the CGI is poor at both sampling ratios, where 15% is better than 5%, as expected. Specifically, the SSIM at 15% is much better as compared to the case of 5%, since at 15%, some of the digits can already be identified by eyes. Again, since the pixel size is small, the performance of CS is even worse than the traditional CGI. The image quality is further degraded when noise is introduced in the bucket signal. On the other hand, the GT results are always much better than that of the CGI and CS. We note here that, there is no significant difference between GT results with and without introduced noise. On the other hand, the GT results are always much better than that of the CGI and CS. We note here that there is no significant difference between GI results with and without introduced noise for the same sampling ratio. This is because the GT scheme uses the attention mechanism to capture the global context information and ignores part of the noise interference. One may also notice that the 5% sampling ratio results are slightly better than 15%. This is partly because the image quality is already saturated at a 5% sampling ratio. When only ten objects are randomly chosen for the compassion, the indicators might have better values at a 5% sampling ratio. However, when the sampling ratio is 15%, the detailed structure of the image can be better retrieved. This can be seen clearly in digits 3, 5, and 7, for instance. That being said, the most significant improvement when increasing the sampling ratio is the accuracy of the GT output, which is an essential metric that generally describes how the model performs [46]. When 100 digits from the testing dataset are used, the accuracy of the GT is only ∼ 30% at 5% sampling ratio, but increases to 91% at 15% sampling ratio. In general, the accuracy can be further improved using more compute such as tokens, training dataset size, etc.

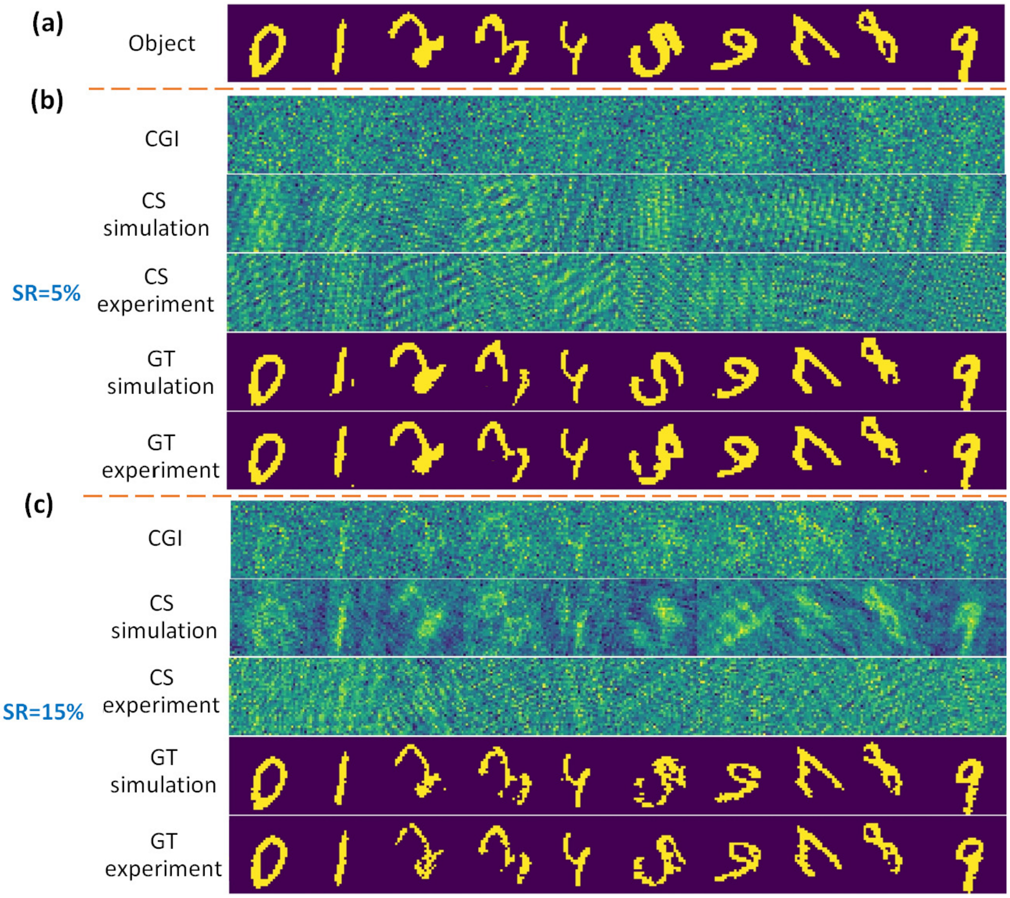

Fig. 4. GT results with Rayleigh speckles at sampling ratio of 5% and 15%. (a) The object, digits 0-9 from the MNIST handwritten dataset. (b) Results with SR=5%. The first row is the traditional CGI. The second and third rows are the CS simulation and experimental results. The fourth and fifth are the GT simulation and experimental results. (c) Same as (b) with SR=15%.

3.3. GT with customized speckles

Next, we show that the GT results can be further improved when customized speckles are used. When standard random speckle patterns interact with the object, the intrinsic connection or correlation between the bucket signals is from the object itself. Our recent work has demonstrated that speckles can be optimized via deep learning at the sub-Nyquist condition [47]. Due to the unique statistic feature, the results outperform the deep learning GI process when the speckles are used for the CGI. The particular statistics of the speckles also suggest that the bucket signals have a unique connection. This type of connection could benefit the attention mechanism used in the

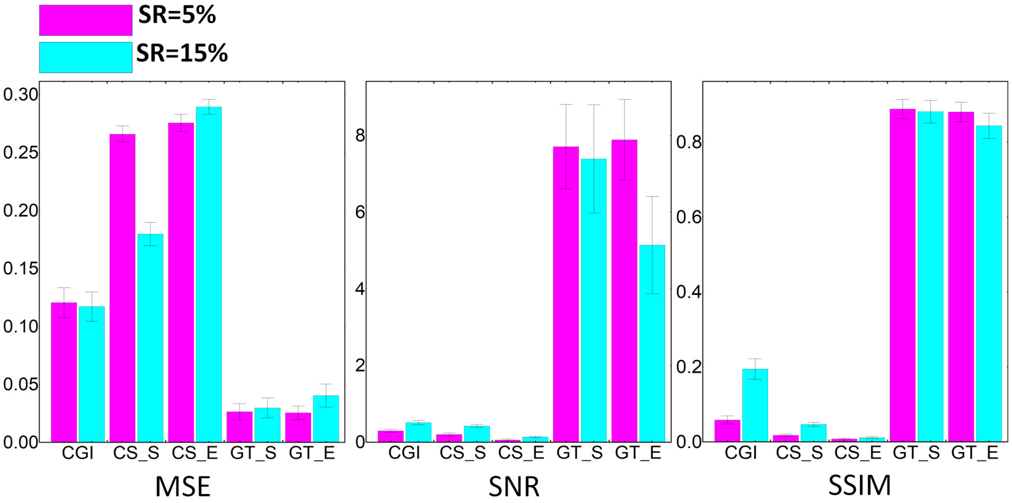

Fig. 5. MSE, SNR, and SSIM of the results shown in Fig. 4. CS_S, CS_E, GT_S, GT_E are the CS simulation and experimental results, GT simulation and experimental results, respectively.

GT network. We, therefore, use these speckle patterns for the GT process for better performance. Here we used three different sampling ratios at 2% (21 speckle pattern), 3% (31 speckle patterns), and 5%. In addition to the MNIST dataset, we used the Quick Draw smiley face dataset to examine the generalization of the GT method. The simulation and experimental results of one typical example each from the two datasets are shown in Fig. 6. It is clearly seen that the GT results can tentatively retrieve both images at the sampling ratio as low as 2%. For sampling ratios of 2%, 3%, and 5%, the overall accuracy is 36%, 66%, and 80%, respectively. We notice that at 5%, the GT results are almost identical to the original objects.

More evidence is shown in Fig. 7, where ten digits and ten randomly selected smiley faces are used. It can be seen that almost all the ten digits and ten smiley faces are retrieved in GT simulation and experiments. Not only there is no background noise in the retrieved images, but the detailed structures are almost recovered from the network. This is a significant difference than when the Rayleigh speckle patterns are used to generate the bucket signals. This could be mainly due to that, the self-attention mechanism highly emphasizes the connection between the encoded inputs. When the input bucket signal are generated with speckle patterns that are random (therefore almost orthogonal to each other), less intrinsic connections are encoded in the bucket signals. On the other hand, due to the unique property of the customized speckle patterns, the bucket signals also contain enough information that can later be recognized by the transformer network. Again, even when noise is introduced in the experiment, the results are unaffected. The results of the other two computational imaging methods, CGI and CS, are also presented in Fig. 7. The background noise is suppressed in general, when customized speckle patterns are used. However, the overall imaging quality is still far worse than the GT results.

The results described by the image quality indicators are depicted in Fig. 8. The MSE, SNR, and SSIM values all suggest that the quality of the digits are better than the smiley faces. This could be the fact that the Quick Draw dataset is more diverse in the structure since it is created casually. However, as indicated by the MSE and SSIM values, the experimental results for both datasets are almost identical to the original objects. Further improvement, especially in regard to the accuracy, may be achieved by increasing the training set size, imaging size, increasing the sampling ratio, or optimising the network, etc.

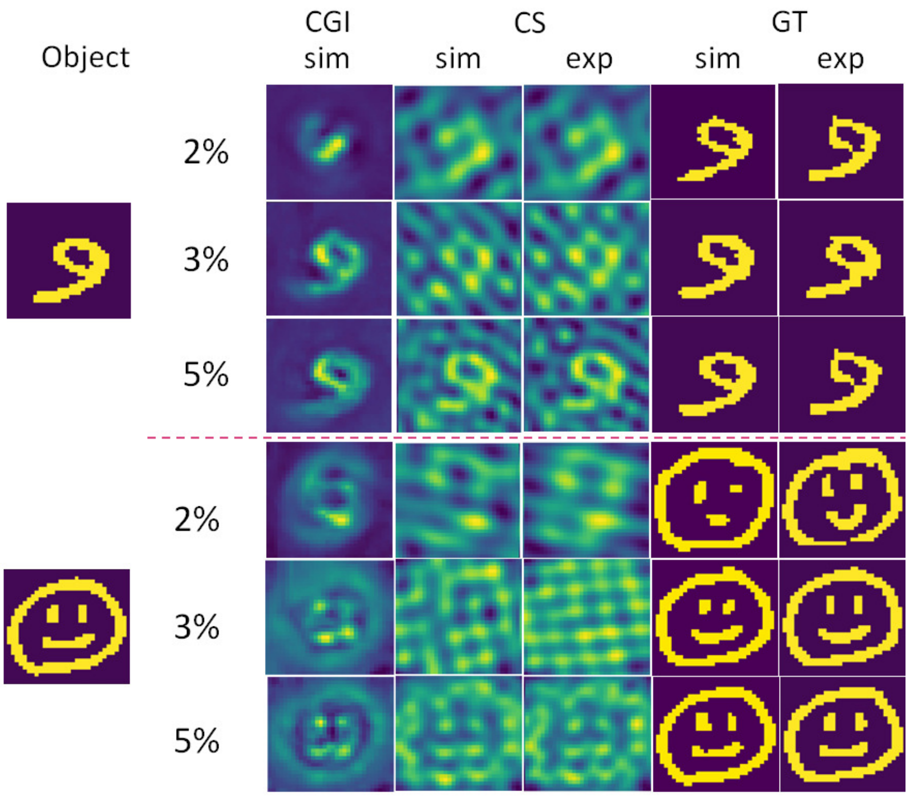

Fig. 6. Results of CGI, CS, and GT with customized speckles. The simulation (sim) and experimental (exp) results of one typical object from each dataset are presented at sampling ratios of 2%, 3%, and 5%, respectively.

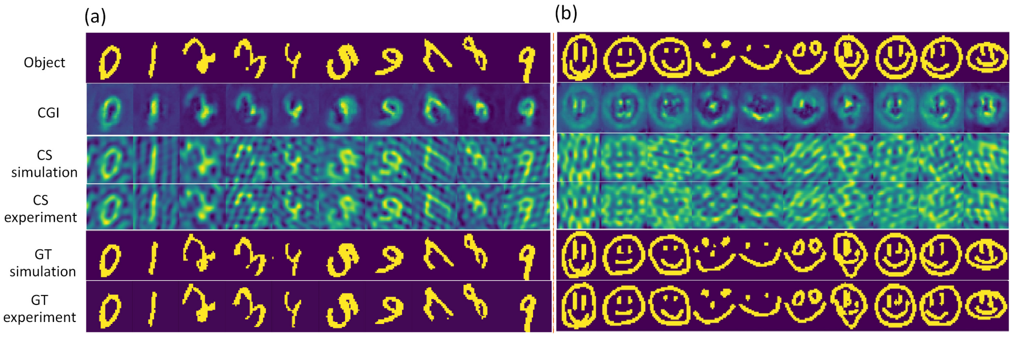

Fig. 7. Comparison of CGI, CS, and GT results of (a)digit numbers 0-9, and (b) ten randomly chosen smiley faces. The sampling ratio of $\beta = 5\%$ is used. The first column is the ground truth object. The second to sixth columns are the CGI simulation results, the CS simulation and experimental results, the GT simulation and experimental results, respectively.

Optics EXPRESS

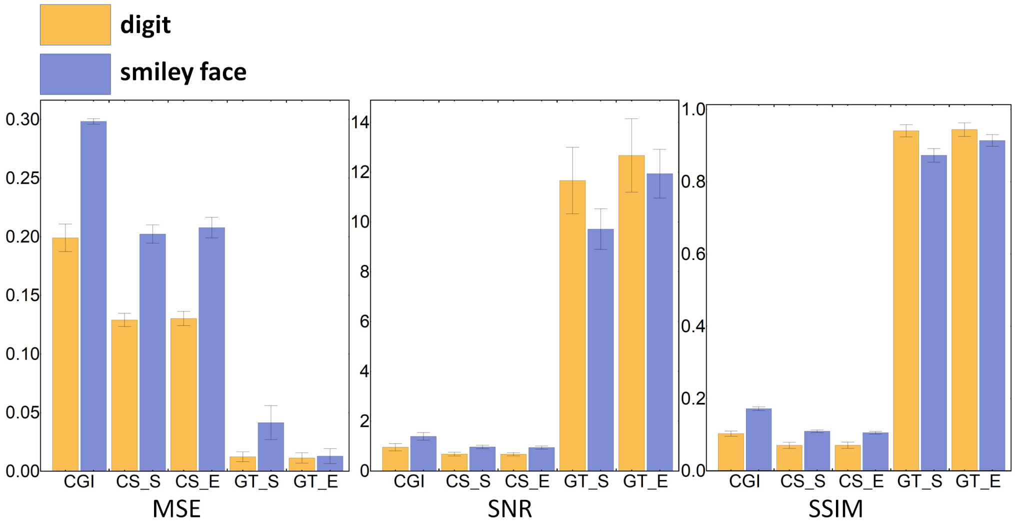

Fig. 8. MSE, SNR, and SSIM of the results shown in Fig. 7.

4. Conclusion

In summary, we have developed and demonstrated a translation mechanism based computational imaging. The object image can be translated only from the bucket detector signals. We have shown that the transformer can be well trained using only simulation data, so the cost of training can be significantly reduced. The proposed method was verified on two different datasets at different sampling ratios through simulations and experiments. Our observation suggests that the proposed method can restore the object image at extremely low sampling ratios. It is also robust to noise interference due to the unique attention mechanism used in the network. This has significant potential to increase the time efficiency of data acquisition in practical applications, especially in noisy environments.

Funding. Air Force Office of Scientific Research (FA9550-20-1-0366 DEF); Office of Naval Research (N00014-20-1-2184); Welch Foundation (A-1261); National Science Foundation (PHY-2013771).

Acknowledgment. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. The authors thank H. T. Song for his helpful discussion.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. S. Cheng, S. Fu, Y. M. Kim, W. Song, Y. Li, Y. Xue, J. Yi, and L. Tian, "Single-cell cytometry via multiplexed fluorescence prediction by label-free reflectance microscopy," Sci. Adv. **7**(3), eabe0431 (2021).
2. J. Peng, M. Yao, J. Cheng, Z. Zhang, S. Li, G. Zheng, and J. Zhong, "Micro-tomography via single-pixel imaging," Opt. Express **26**(24), 31094–31105 (2018).
3. A. M. Kingston, D. Pelliccia, A. Rack, M. P. Olbinado, Y. Cheng, G. R. Myers, and D. M. Paganin, "Ghost tomography," Optica **5**(12), 1516–1520 (2018).
4. P. Gattinger, J. Kilgus, I. Zorin, G. Langer, R. Nikzad-Langerodi, C. Rankl, M. Gröschl, and M. Brandstetter, "Broadband near-infrared hyperspectral single pixel imaging for chemical characterization," Opt. Express **27**(9), 12666–12672 (2019).
5. B. Zeng, Z. Huang, A. Singh, Y. Yao, A. K. Azad, A. D. Mohite, A. J. Taylor, D. R. Smith, and H.-T. Chen, "Hybrid graphene metasurfaces for high-speed mid-infrared light modulation and single-pixel imaging," Light: Sci. Appl. **7**(1), 51 (2018).
6. R. I. Stantchev, B. Sun, S. M. Hornett, P. A. Hobson, G. M. Gibson, M. J. Padgett, and E. Hendry, "Noninvasive, near-field terahertz imaging of hidden objects using a single-pixel detector," Sci. Adv. **2**(6), e1600190 (2016).

7. J. H. Shapiro, “Computational ghost imaging,” Phys. Rev. A **78**(6), 061802 (2008).
8. Y. Bromberg, O. Katz, and Y. Silberberg, “Ghost imaging with a single detector,” Phys. Rev. A **79**(5), 053840 (2009).
9. T. B. Pittman, Y. H. Shih, D. V. Strekalov, and A. V. Sergienko, “Optical imaging by means of two-photon quantum entanglement,” Phys. Rev. A **52**(5), R3429–R3432 (1995).
10. A. Valencia, G. Scarcelli, M. D’Angelo, and Y. Shih, “Two-photon imaging with thermal light,” Phys. Rev. Lett. **94**(6), 063601 (2005).
11. X.-H. Chen, Q. Liu, K.-H. Luo, and L.-A. Wu, “Lensless ghost imaging with true thermal light,” Opt. Lett. **34**(5), 695–697 (2009).
12. M.-J. Sun, L.-T. Meng, M. P. Edgar, M. J. Padgett, and N. Radwell, “A russian dolls ordering of the hadamard basis for compressive single-pixel imaging,” Sci. Rep. **7**(1), 3464 (2017).
13. Z. Zhang, X. Wang, G. Zheng, and J. Zhong, “Hadamard single-pixel imaging versus fourier single-pixel imaging,” Opt. Express **25**(16), 19619–19639 (2017).
14. P. Stockton, G. Murray, J. J. Field, J. Squier, A. Pezeshki, and R. A. Bartels, “Tomographic single pixel spatial frequency projection imaging,” Opt. Commun. **520**, 128401 (2022).
15. X. Zhao, X. Nie, Z. Yi, T. Peng, and M. O. Scully, “Imaging through scattering media via spatial–temporal encoded pattern illumination,” Photonics Res. **10**(7), 1689–1694 (2022).
16. O. Katz, Y. Bromberg, and Y. Silberberg, “Compressive ghost imaging,” Appl. Phys. Lett. **95**(13), 131110 (2009).
17. V. Katkovnik and J. Astola, “Compressive sensing computational ghost imaging,” J. Opt. Soc. Am. A **29**(8), 1556–1567 (2012).
18. X. Nie, X. Zhao, T. Peng, and M. O. Scully, “Sub-nyquist computational ghost imaging with orthonormal spectrum-encoded speckle patterns,” Phys. Rev. A **105**(4), 043525 (2022).
19. M. Cao, X. Yang, J. Wang, S. Qiu, D. Wei, H. Gao, and F. Li, “Resolution enhancement of ghost imaging in atom vapor,” Opt. Lett. **41**(22), 5349–5352 (2016).
20. N. Bender, M. Sun, H. Yılmaz, J. Bewersdorf, and H. Cao, “Circumventing the optical diffraction limit with customized speckles,” Optica **8**(2), 122–129 (2021).
21. D. Pelliccia, A. Rack, M. Scheel, V. Cantelli, and D. M. Paganin, “Experimental x-ray ghost imaging,” Phys. Rev. Lett. **117**(11), 113902 (2016).
22. L. Olivieri, J. S. T. Gongora, L. Peters, V. Cecconi, A. Cutrona, J. Tunesi, R. Tucker, A. Pasquazi, and M. Peccianti, “Hyperspectral terahertz microscopy via nonlinear ghost imaging,” Optica **7**(2), 186–191 (2020).
23. R. I. Khakimov, B. Henson, D. Shin, S. Hodgman, R. Dall, K. Baldwin, and A. Truscott, “Ghost imaging with atoms,” Nature **540**(7631), 100–103 (2016).
24. A. Trimeche, C. Lopez, D. Comparat, and Y. Picard, “Ion and electron ghost imaging,” Phys. Rev. Res. **2**(4), 043295 (2020).
25. F. Wang, H. Wang, H. Wang, G. Li, and G. Situ, “Learning from simulation: An end-to-end deep-learning approach for computational ghost imaging,” Opt. Express **27**(18), 25560–25572 (2019).
26. S. Rizvi, J. Cao, K. Zhang, and Q. Hao, “Deepghost: real-time computational ghost imaging via deep learning,” Sci. Rep. **10**(1), 11400 (2020).
27. H. Wu, R. Wang, G. Zhao, H. Xiao, J. Liang, D. Wang, X. Tian, L. Cheng, and X. Zhang, “Deep-learning denoising computational ghost imaging,” Opt. Lasers Eng. **134**, 106183 (2020).
28. J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *European conference on computer vision*, (Springer, 2016), pp. 694–711.
29. G. Barbastathis, A. Ozcan, and G. Situ, “On the use of deep learning for computational imaging,” Optica **6**(8), 921–943 (2019).
30. M. Lyu, W. Wang, H. Wang, H. Wang, G. Li, N. Chen, and G. Situ, “Deep-learning-based ghost imaging,” Sci. Rep. **7**(1), 17865 (2017).
31. H. Song, X. Nie, H. Su, H. Chen, Y. Zhou, X. Zhao, T. Peng, and M. O. Scully, “0.8% nyquist computational ghost imaging via non-experimental deep learning,” Opt. Commun. **520**, 128450 (2022).
32. R. Collobert and J. Weston, “A unified architecture for natural language processing: Deep neural networks with multitask learning,” in *Proceedings of the 25th international conference on Machine learning*, (2008), pp. 160–167.
33. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems **30** (2017).
34. T. Wolf, L. Debut, and V. Sanh, *et al.*, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, (2020), pp. 38–45.
35. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 10012–10022.
36. H. Chen, T. Peng, and Y. Shih, “100% correlation of chaotic thermal light,” Phys. Rev. A **88**(2), 023808 (2013).
37. E. J. Candes, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” Comm. Pure Appl. Math. **59**(8), 1207–1223 (2006).
38. W. Shi, F. Jiang, S. Zhang, and D. Zhao, “Deep networks for compressed image sensing,” in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, (IEEE, 2017), pp. 877–882.

Optics EXPRESS

39. J. Zhang and B. Ghanem, "Ista-net: Interpretable optimization-inspired deep network for image compressive sensing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 1828–1837.
40. S. Jiao, Y. Gao, J. Feng, T. Lei, and X. Yuan, "Does deep learning always outperform simple linear regression in optical imaging?" Opt. Express **28**(3), 3717–3731 (2020).
41. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT press, 2016).
42. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," Advances in neural information processing systems **27** (2014).
43. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078 (2014).
44. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation **9**(8), 1735–1780 (1997).
45. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. on Image Process. **13**(4), 600–612 (2004).
46. D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp. 3163–3172.
47. X. Nie, H. Song, W. Ren, X. Zhao, Z. Zhang, T. Peng, and M. O. Scully, "Deep-learned speckle pattern and its application to ghost imaging," arXiv, arXiv:2112.13293 (2021).10.48550/arXiv.2112.13293