Chinese Journal of Electronics
Vol.32, No.5, Sept. 2023

Transformer-Based Under-sampled Single-Pixel Imaging

TIAN Ye[1,4], FU Ying[2,3], and ZHANG Jun[1,2,4]

(1. *School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China*)

(2. *Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing 314019, China*)

(3. *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*)

(4. *Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China*)

Abstract — Single-pixel imaging, as an innovative imaging technique, has attracted much attention during the last decades. However, it is still a challenging task for single-pixel imaging to reconstruct high-quality images with fewer measurements. Recently, deep learning techniques have shown great potential in single-pixel imaging especially for under-sampling cases. Despite outperforming traditional model-based methods, the existing deep learning-based methods usually utilize fully convolutional networks to model the imaging process which have limitations in long-range dependencies capturing, leading to limited reconstruction performance. In this paper, we present a transformer-based single-pixel imaging method to realize high-quality image reconstruction in under-sampled situation. By taking advantage of self-attention mechanism, the proposed method is good at modeling the imaging process and directly reconstructs high-quality images from the measured one-dimensional light intensity sequence. Numerical simulations and real optical experiments demonstrate that the proposed method outperforms the state-of-the-art single-pixel imaging methods in terms of reconstruction performance and noise robustness.

Key words — Computational imaging, Single-pixel imaging, Vision transformer, Under-sampled ratio.

I. Introduction

Single-pixel imaging (SPI) [1]–[3], as a novel computational imaging technique, has attracted a wide range of attention recently due to its ability of replacing conventional pixel-rich detectors with a single-pixel detector. Different from the conventional single-shot imaging scheme, single-pixel imaging reconstructs object image with multiple measurements. To be more specific, SPI uses a sequence of modulation light patterns to illuminate the object image and the corresponding reflected or transmitted light is captured by a single-element photodetector as one-dimensional measurement data. The object image can be recovered from the recorded measurements by using various computational imaging algorithms. SPI has unique advantages such as high signal-to-noise ratio, low cost, broadband, and flexible light-path configuration [4]. Therefore, it has been applied to three-dimensional imaging [5], gas imaging [6], terahertz imaging [7], remote sensing [8], and many other fields [9] where pixelated detectors are not accessible due to cost or technical constraints.

In SPI, high-quality image reconstruction needs a large amount of measurements, which greatly increases the imaging time and limits the practical application. Therefore, it is very important for SPI to balance imaging quality and efficiency. Various methods have been proposed to solve this problem. One of the most commonly used methods is to adopt the compressed sensing (CS) which incorporates the prior knowledge that most natural images have [10], [11]. However, these hand-crafted priors are insufficient to represent the diverse range of real-world images. Moreover, these CS-based methods are very time-consuming because of the iterative process. Recently, deep learning methods have been proposed for SPI to improve the reconstructed image quality [12]–[14]. In comparison to the CS-based methods, deep learning (DL)-based methods can recover higher quality images with fewer measurements. But most of these methods need data preprocessing to recover the approximant first, which ignores the SPI physical model and is more like image denoising. More importantly, for SPI, the measured data has a strong

Manuscript Received Aug. 22, 2022; Accepted Jan. 18, 2023. This work was supported by the National Key R&D Program of China (2022YFC3300704) and the National Natural Science Foundation of China (62171038, 61827901, 62088101).

DOI:10.23919/cje.2022.00.284

relationship because it is obtained from the same scene, indicating that long-range dependencies are crucial for effective SPI modeling. But the current DL-based methods all rely on convolution filters to model the dependencies across the input and desirable reconstructed image. These convolutional neural network (CNN) based models often focus on local features and ignore the long-range dependencies of SPI measurements. Thus, these DL-based methods exhibit satisfactory reconstruction results only for simple objects and show little performance improvement as the number of measurements increases.

Inspired by the great success in the field of natural language processing (NLP) [15], transformer models are applied to computer vision (CV) tasks recently and show valuable potential to substitute for CNN models [16], [17]. Different from CNN models, transformer models are good at extracting global features and flexibly modeling long-range dependencies of data at various scales. Benefiting from these advantages, transformer models have the potential to overcome the limitations of CNN-based models in SPI reconstruction.

In this paper, we propose an under-sampled transformer-based SPI method, which can realize high-quality image reconstruction with the one-dimensional light intensity sequence. By developing an end-to-end transformer network, the long-range dependencies in the measured data can be effectively utilized to improve the reconstruction performance of under-sampled SPI. Simulated and experimental results show that our method can achieve better performance than the state-of-the-art SPI methods in terms of fidelity and robustness. This work provides a novel solution for high-quality under-sampled single-pixel imaging.

II. Related Work

In this section, we review the researches most relevant to our work, including single-pixel imaging and vision transformer. Through this review, the innovation of our proposed method is clarified.

1. Single-pixel imaging

Existing SPI methods can be divided into conventional model-based methods and DL-based methods. Among them, the conventional model-based SPI methods can be classified into two categories: non-iterative SPI methods and iterative SPI methods. The non-iterative SPI methods [18] directly utilize the correlation between modulation light patterns and object image to reconstruct the object image without iteration, such as the differential ghost imaging (DGI) method [19]. However, these non-iterative SPI methods can successfully reconstruct images only in the fully-sampled situation. The iterative SPI methods, including gradient descent SPI methods [20], alternating projection SPI methods [21] and CS-based SPI methods [10], [11], combine convex optimization theory and various of hand-crafted priors to reconstruct the object images. In [20], the authors compared all model-based SPI methods and demonstrated that the CS-based methods outperform other model-based methods in under-sampled cases. Nevertheless, the CS-based methods are with higher algorithm complexity and not suitable for real-time imaging.

Recently, DL-based SPI methods are widely concerned due to their superior SPI reconstruction performance in the under-sampled situation. In [12], researchers reported a DL-based ghost imaging method (GIDL) with a two-step process. It first uses a traditional correlation-based method to reconstruct the approximate image from the under-sampled measurements, then uses a deep neural network (DNN) to improve the reconstruction performance. Similarly, in [13], researchers proposed a deep learning ghost imaging method (DLGI). Different from the GIDL, it uses a CS-based method in the first step and a CNN-based network in the second step. However, these methods reduce the reconstruction efficiency due to the long time consumed in the first step. To improve the reconstruction efficiency, in [22], researchers proposed a one-step SPI method based on a deep convolutional auto-encoder network (DCAN), which can reconstruct image directly from the under-sampled SPI measurements. After that, researchers proposed several other one-step SPI methods that employ different end-to-end SPI reconstruct networks [23]–[26]. However, these CNN-based network structures are inefficient in reasoning long-range dependencies in SPI measurements. Therefore, these methods are more suitable for MNIST-like simple object imaging. To further enhance the reconstruction performance, researchers prefer making efforts to improve the two-step DL-based methods recently [27]–[29]. These methods use non-iterative SPI methods to reconstruct the approximate image in the first step and more complicated CNN-based networks in the second step, which can obtain better reconstruction efficiency and quality compared with the previous two-step methods. Besides, researchers applied a physics-enhanced framework with fine-tuning process [30] to improve the reconstruction quality and generalization of the two-step method [31]. In summary, the majority of DL-based SPI methods adopt a two-step process for imaging [26]–[31] and heavily rely on CNN structures for modeling.

2. Vision transformer

Transformer was first proposed in [32] and has gained extensive application for NLP tasks [33]. The key component of transformer is attention mechanism, which can capture long-term information between sequence elements. Recently, many efforts have been made by researchers to explore its applicability in CV tasks. Compared with CNN-based architecture, transformer shows more appealing performance in various applications, such as image classification [34], segmentation [35], object detection [36], and human pose estimation [37]. Among them, ViT [38] is the first work that uses transformer in place of the standard convolution. To adapt to visual tasks, the 2D image patches are converted into a vector and fed into the transformer. After that, many kinds of transformers are developed for different demands. For example, to further reduce computation expense and improve the efficiency of transformer, pyramid vision transformer (PVT) was proposed in [39], which makes full use of spatial-reduction attention (SRA) to learn multiscale and high-resolution features. To improve the modeling capacity of local information, shifted windows (Swin) transformer was proposed in [40] which has the advantage of processing large-size images on the strength of shifted windows mechanism. In summary, transformer now is becoming an upgraded alternative for original CNNs in CV tasks due to its outstanding performance.

III. Principles and Methods

In this section, we first formulate the problem for under-sampled SPI reconstruction and illustrate the motivation of our work. Then we describe the proposed transformer-based single-pixel imaging method in detail.

1. Problem formulation

The process of single-pixel imaging consists of two stages. First, the object is illuminated by a series of modulation light patterns and the corresponding reflected light is collected as measurement data with a single-pixel detector. Mathematically, this process can be expressed as

$$B_i = \int P_i(x, y)T(x, y)\mathrm{d}x\mathrm{d}y \tag{1}$$

where $T(x, y)$ denotes the object image and the total number of pixels is N, (x, y) is the transverse coordinates at the object plane. $P_i(x, y)$ denotes the i-th modulation pattern, where $i = 1, 2, \ldots, M$ and M is the total number of modulation patterns. In under-sampled SPI condition, N is larger than M and M/N is the sampling ratio. B_i is the reflected light intensity under the i-th modulation pattern. For convenience, we generally express the above process in matrix form as

$$\boldsymbol{B} = \boldsymbol{PT} \tag{2}$$

Secondly, various kinds of SPI algorithms are used to reconstruct the object image $\boldsymbol{T}$ according to the known modulation light patterns $\boldsymbol{P}$. In the under-sampled situation, the reconstruction of object image is an ill-posed problem. For CS-based SPI methods, the image reconstruction is regarded as an optimization problem, which can be expressed as

$$\hat{\boldsymbol{T}} = \arg\min_{\boldsymbol{T}} \|\boldsymbol{B} - \boldsymbol{PT}\|_2^2 + \tau R(\boldsymbol{T}) \tag{3}$$

where $R(\boldsymbol{T})$ is the prior that most natural images possess, such as sparsity, total variation and low rankness. τ is a trade-off parameter. By combining the convex optimization theory [41], [42], the above optimization problem can be solved and the object image can be obtained. However, these CS-based methods need to tune parameters manually and have slow reconstruction speed due to their iterative process. By comparison, DL-based SPI methods adopt a CNN-based network that can implicitly learn the prior to reconstruct the object images. Most of these methods need preprocessing algorithms [12], [13] to recover the approximant first from measurements $\boldsymbol{B}$, and then send the noisy image to the CNN-based network to get a higher quality reconstructed image. This process can be expressed as

$$\hat{\boldsymbol{T}} = f_{\text{cnn}}(\text{SPI}(\boldsymbol{B})) \tag{4}$$

where $f_{\text{cnn}}(\cdot)$ denotes the CNN-based SPI reconstruction network. $\text{SPI}(\cdot)$ denotes the preprocessed algorithm, such as DGI algorithm [19]. However, due to the limitation of CNN structure, these CNN-based SPI methods perform well only in reconstructing binary images and sparse gray images. Besides, when the number of measurements increases, the performance improvement of these methods is very limited. It is still a challenging task for SPI to reconstruct high-quality images directly from fewer one-dimensional measurement data. Therefore, we develop a transformed-based single-pixel imaging method to reconstruct the high-quality image directly from one-dimensional light intensity sequence in the under-sampled situation.

2. Transformer-based single-pixel imaging

The proposed method employs a transformer-based network to reconstruct the object image directly from the one-dimensional measurements $\boldsymbol{B}$. This reconstruction process can be mathematically expressed as an implicit function:

$$\hat{\boldsymbol{T}} = f_{\text{tspi}}(\boldsymbol{B}) \tag{5}$$

where $f_{\text{tspi}}(\cdot)$ denotes our proposed transformed-based SPI network that models the dependencies across the SPI measurements to the reconstructed image. $\hat{\boldsymbol{T}}$ denotes the reconstructed object image. The mapping from one-dimensional measurements to a two-dimensional image without knowing the transformation matrix is highly ill-posed. Therefore, we train the proposed transformer-based network from K pairs of labeled data each of which pairs up a known object image $\boldsymbol{T}^k$ and the corresponding SPI measurements $\boldsymbol{B}^k$, where $k=1, 2, \ldots, K$. The training stage can be expressed as:

$$\tilde{f}_{\text{tspi}} = \arg\min_{w} \sum_{k=1}^{K} L(\boldsymbol{T}^k, f_{\text{tspi},w}(\boldsymbol{B}^k)) \tag{6}$$

where w is the all learnable parameters in our proposed network, $L(\cdot)$ is a loss function to measure the error between the network output $f_{\text{tspi}}(\boldsymbol{B}^k)$ and the ground-truth $\boldsymbol{T}^k$. We use the mean squared error (MSE) as the loss in this work. After training is completed, the arbitrary object image can be reconstructed in terms of its SPI measurements:

$$\hat{\boldsymbol{T}}' = \tilde{f}_{\text{tspi}}(\boldsymbol{B}') \tag{7}$$

Then, we illustrate the architecture of the proposed transformed-based SPI network in Fig.1. As shown in Fig.1, the input of the network is the one-dimensional measurement data obtained by the single-pixel detector. To effectively exploit the long-range dependencies in measurements, we first add a fully connected layer at the input which reshapes the measurement data into a two-dimensional feature map. Then, we use a convolutional layer to extract the low-level features. Next, we use four transformer blocks and one convolutional layer to extract the deep features. Finally, we adopt one convolution layer to reconstruct the object image from the extracted features. In order to obtain better image quality, we adopt a long skip connection that can aggregate both low-level features and deep features to the last convolution layer. For more details, each transformer block consists of four transformer layers and one convolutional layer. Inspired by the advantages of the Swin transformer, each transformer layer consists of LayerNorm (LN), multi-layer perceptron (MLP) and multi-head self-attention (MSA). In addition, the shifted window mechanism is used in the MSA, which is expressed as S-MSA. For the j-th transformer layer, assuming $\boldsymbol{X}_{j-1}$ is the input, the output $\boldsymbol{X}_j$ can be expressed as:

$$\boldsymbol{X}_j' = \text{S-MSA}(\text{LN}(\boldsymbol{X}_{j-1})) + \boldsymbol{X}_{j-1} \tag{8}$$

$$\boldsymbol{X}_j = \text{MLP}(\text{LN}(\boldsymbol{X}_j')) + \boldsymbol{X}_j' \tag{9}$$

where $\boldsymbol{X}_j'$ denotes the output of the S-MSA.

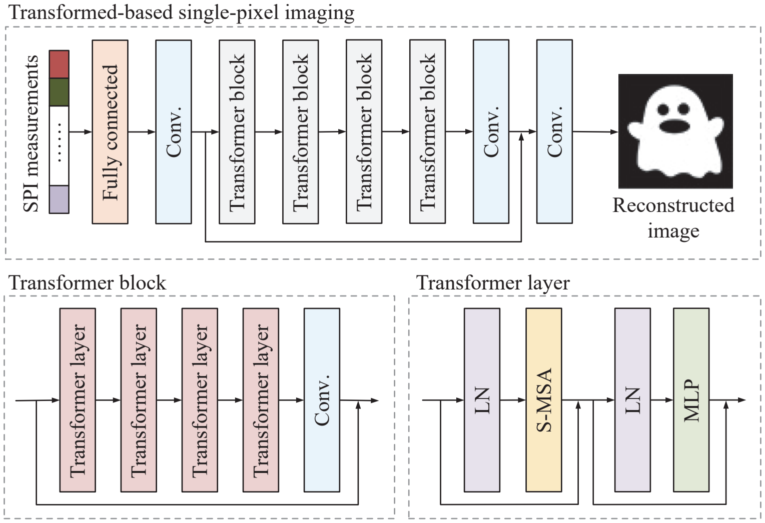


Fig. 1. Architecture of the proposed transformer-based SPI network.

IV. Simulation and Experimental Results

In this section, we conduct several simulations and real experiments to evaluate the performance of our method. We first describe the settings for our simulations and real experiments, including dataset, quantitative evaluation metrics and competing methods. Then, we compare our proposed method with the state-of-the-art methods by simulation, in which both the noiseless and noisy situations are considered. Finally, we implement our proposed method on the real SPI captured data which further verifies the effectiveness of our proposed method.

1. Metrics and setups

We conduct a comparative analysis between our proposed transformer-based SPI method and the state-of-the-art SPI methods, which includes traditional CS-based methods (i.e., CS-sparse [10] and CS-TV [11]), and DL-based methods (i.e., DCAN [22], RNN [25], and physics-enhanced method [31]). Following [22], the optimized binary patterns are selected as the modulation light patterns because they are more practical and hardware-friendly. More specifically, the values of modulation patterns are optimized together with the proposed network in the training stage and restricted to $+1$ or -1 approximately by regularization function. We use Adam optimizer [43] to train the proposed transformer-based SPI network. The learning rate is 1×10^{-4}. The batch size is set to 16 with 150 epochs. All DL-based methods are trained using STL-10 dataset [22] on a server equipped with a GeForce RTX 3090. All competing methods are tested on a computer with NVIDIA GeForce GTX 1660 SUPER GPU, 16 GB RAM, and 64 bit Windows 10 operating system. To quantitatively evaluate all methods, we employ two performance metrics, including peak signal-to-noise ratio (PSNR) and structure similarity (SSIM) index [44]. In general, the larger the PSNR and SSIM values, the better the reconstruction performance of SPI.

2. Simulations

We compare the performance of all competing SPI algorithms based on Set12 [45] dataset. The sampling ratios are set as 5%, 8%, 10%, 20%, and 30%. The quantitative results of SPI reconstruction are shown in Table 1. The proposed transformer-based SPI method consistently outperforms both CS-based and DL-based SPI methods in terms of PSNR and SSIM, demonstrating superior reconstruction quality across various sampling ratios.

To facilitate a visual comparison of all competing methods, we present the reconstructed images of "cameraman" at various sampling ratios for each competing methods, as shown in Fig.2. The image have 64×64 pixels. From Fig.2, it is apparent that more details and sharper edges can be obtained by our proposed method, which further demonstrates the effectiveness of our proposed method.

In practical SPI, the measurements inevitably contain various noise. To test the noise robustness of our method, we simulate the case that Gaussian white noise is included in SPI measurements. Following [20], the noise level added in SPI measurements can be calculated by the dividing of noise standard deviation and total pixel number. According to the above definition, we set the noise level as 1×10^{-5} to 1×10^{-3}. The im-

Table 1. Average PSNR (dB) and SSIM of the reconstructed results at various sampling ratios on Set12 dataset

Image size: 32 × 32							
Algorithm	Metrics	Sampling ratio					
		2%	5%	8%	10%	20%	30%
CS-sparse	PSNR	11.80	12.54	12.82	13.05	14.73	16.21
	SSIM	0.0527	0.0824	0.1233	0.1480	0.3194	0.4501
CS-TV	PSNR	13.83	15.03	15.80	16.23	18.05	19.38
	SSIM	0.1776	0.2672	0.3464	0.3907	0.5551	0.6522
DCAN	PSNR	17.85	19.16	19.91	20.98	22.78	22.45
	SSIM	0.2662	0.4149	0.4945	0.5586	0.7112	0.6848
RNN	PSNR	18.00	19.82	20.69	21.09	23.20	23.08
	SSIM	0.3085	0.5040	0.5934	0.6307	0.7822	0.7830
Physics-enhanced	PSNR	16.78	18.23	19.27	19.97	22.26	23.50
	SSIM	0.2735	0.4211	0.5068	0.5426	0.7074	0.7693
Ours	PSNR	**18.72**	**20.60**	**21.55**	**22.40**	**23.64**	**25.04**
	SSIM	**0.3350**	**0.5322**	**0.6145**	**0.6811**	**0.7853**	**0.8301**

Image size: 64 × 64							
Algorithm	Metrics	Sampling ratio					
		2%	5%	8%	10%	20%	30%
CS-sparse	PSNR	12.24	13.24	14.02	14.54	16.51	18.12
	SSIM	0.0554	0.1127	0.1525	0.1882	0.3440	0.4661
CS-TV	PSNR	15.17	16.60	17.51	17.93	19.89	21.39
	SSIM	0.2584	0.3470	0.4138	0.4478	0.5807	0.6662
DCAN	PSNR	18.77	19.97	20.12	20.68	21.00	21.97
	SSIM	0.3531	0.4720	0.4816	0.5305	0.5449	0.6308
RNN	PSNR	16.53	20.86	21.69	21.52	21.45	23.16
	SSIM	0.3504	0.5819	0.6461	0.6372	0.6819	0.7428
Physics-enhanced	PSNR	16.85	19.41	20.73	21.36	23.20	24.74
	SSIM	0.3233	0.4678	0.5614	0.6021	0.7052	0.7752
Ours	PSNR	**19.57**	**21.29**	**22.00**	**22.49**	**24.30**	**25.34**
	SSIM	**0.4366**	**0.5928**	**0.6532**	**0.6785**	**0.7870**	**0.8280**

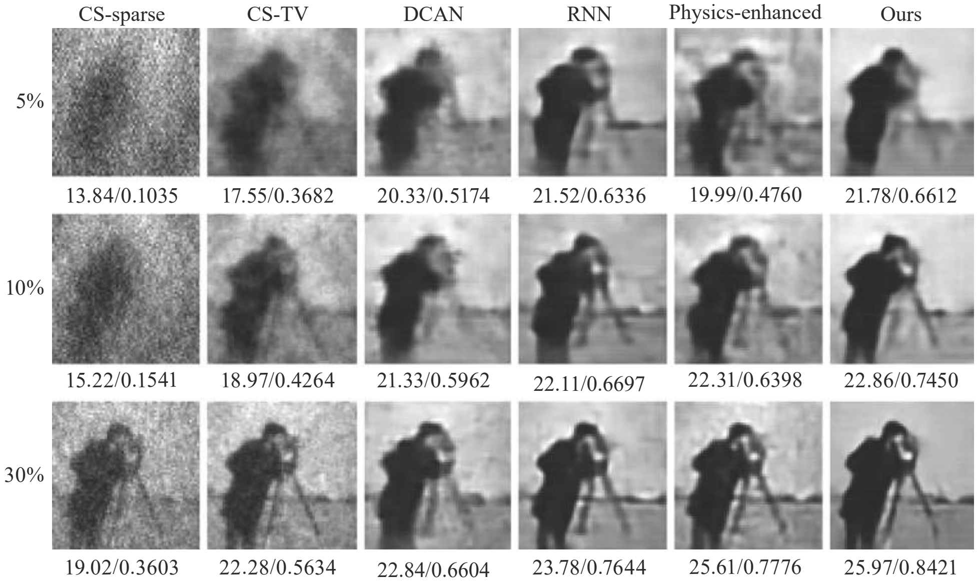


Fig. 2. Reconstructed images of "cameraman" (PSNR(dB)/SSIM) with different methods.

age have 64 × 64 pixels. The sampling ratio is 8%. The simulated results are presented in Table 2. It is obvious that our method maintains the best SPI reconstruction performance, even the noise level increases. It indicates our method has superior noise robustness and much suitable for practical SPI applications.

Table 2. Average PSNR (dB) and SSIM of the reconstructed results on Set12 dataset at different noise levels

Algorithm	Metrics	Noise level				
		1E−3	5E−4	1E−4	5E−5	1E−5
CS-sparse	PSNR	12.99	13.79	14.02	14.02	14.02
	SSIM	0.1100	0.1393	0.1503	0.1523	0.1525
CS-TV	PSNR	16.33	17.15	17.48	17.50	17.51
	SSIM	0.3348	0.3874	0.4122	0.4133	0.4138
DCAN	PSNR	19.86	20.04	20.11	20.12	20.12
	SSIM	0.4710	0.4780	0.4813	0.4815	0.4816
RNN	PSNR	21.19	21.54	21.67	21.68	21.69
	SSIM	0.6188	0.6384	0.6455	0.6459	0.6461
Physics-enhanced	PSNR	19.54	20.40	20.72	20.73	20.73
	SSIM	0.4953	0.5404	0.5571	0.5577	0.5578
Ours	PSNR	**21.31**	**21.83**	**21.99**	**22.00**	**22.00**
	SSIM	**0.6190**	**0.6457**	**0.6530**	**0.6532**	**0.6532**

Additionally, we conduct a comparison of computational costs associated with all competing DL-based SPI methods, where the images have 64 × 64 pixels with 10% sampling ratio. From Table 3, it can be seen that the one-step DL-based SPI methods consume shorter inference time than the two-step method, indicating a higher reconstruction efficiency. Among the one-step methods, our proposed method has a comparable inference time but higher training consumption. However, the relatively high training consumption is acceptable because it does not affect the reconstruction efficiency. In addition, our method has a few more network parameters than DCAN and RNN methods but realizes better reconstruction performance than other competing methods.

Table 3. The comparison of computational cost among different DL-based SPI methods

Algorithm	Params.	Training consumption	Inference time
DCAN	3.37 M	2 hours single-GPU	0.1031 s
		(1545 MiB)	
RNN	5.61 M	27 hours single-GPU	0.1485 s
		(1106 MiB)	
Physics-enhanced	12.87 M	64 hours single-GPU	7.9219 s
		(2093 MiB)	
Ours	6.50 M	59 hours single-GPU	0.1401 s
		(15723 MiB)	

3. Real experiments

To further validate the effectiveness of the transformer-based SPI method, we build up an optical system to acquire real SPI measurement data. The experimental setup is shown in Fig.3. Specifically, we generate the 64 × 64 binary modulation patterns using a projector (Panasonic, X416C XGA) and project them onto the object. Then we capture the reflection light using a Si amplified photodetector (Thorlabs, PDA100A2). We set the sampling ratio as 5%, 10%, and 30%. Fig.4 illustrates the optical reconstruction results. It is obvious that our method outperforms the model-based methods and DL-based methods, producing higher quality image reconstructions. These experimental results are in good agreement with the simulation results. Furthermore, these results demonstrate that our method is better suited for practical applications.

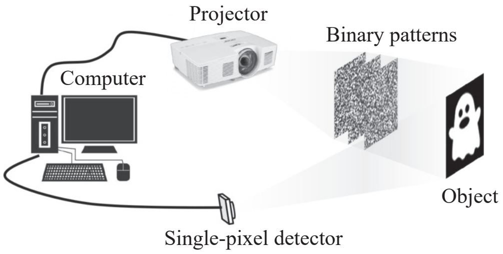


Fig. 3. The SPI experimental setup.

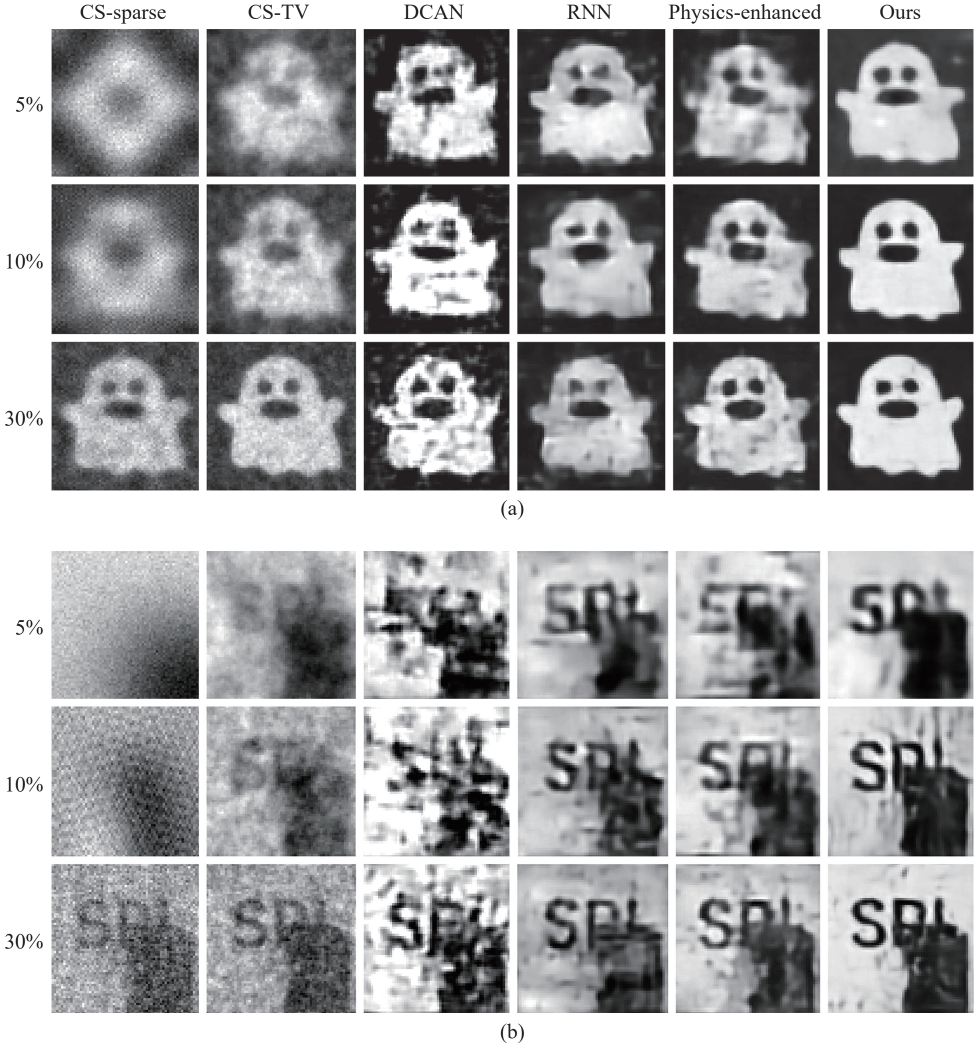


Fig. 4. The images reconstructed by all competing methods using the real SPI system. (a) Imaging of a "ghost" picture; (b) Imaging of a tea caddy with a "SPI" letter background.

V. Conclusions

In this study, we present a novel one-step single-pixel imaging method that enables high-performance SPI reconstruction based on one-dimensional under-sampled measurements. By taking advantage of the self-attention mechanism and shifted window mechanism, the proposed transformer-based SPI network can well exploit the intrinsic features of SPI. Numerous simulation and experimental results show that the proposed method achieves higher image quality and stronger noise robustness than the existing SPI methods. In addition, our work indicates that the transformer architecture has great potential to replace the CNN architecture in single-pixel imaging, which could provide a new insight into optical computational imaging.

References

[1] M. P. Edgar, G. M. Gibson, and M. J. Padgett, "Principles and prospects for single-pixel imaging," *Nature Photonics*, vol.13, no.1, pp.13–20, 2019.

[2] G. M. Gibson, S. D. Johnson, and M. J. Padgett, "Single-pixel imaging 12 years on: A review," *Optics Express*, vol.28, no.19, pp.28190–28208, 2020.

[3] Q. H. Dai, J. M. Wu, J. T. Fan, *et al.*, "Recent advances in computational photography," *Chinese Journal of Electronics*, vol.28, no.1, pp.1–5, 2019.

[4] Y. Tian, Y. Fu, and J. Zhang, "Plug-and-play algorithm for under-sampling Fourier single-pixel imaging," *Science China Information Sciences*, vol.65, no.10, article no.209303, 2022.

[5] M. J. Sun, M. P. Edgar, G. M. Gibson, *et al.*, "Single-pixel three-dimensional imaging with time-based depth resolution," *Nature Communications*, vol.7, article no.12010, 2016.

[6] G. M. Gibson, B. Q. Sun, M. P. Edgar, *et al.*, "Real-time imaging of methane gas leaks using a single-pixel camera," *Optics Express*, vol.25, no.4, pp.2998–3005, 2017.

[7] L. Zanotto, R. Piccoli, J. L. Dong, *et al.*, "Single-pixel terahertz imaging: A review," *Opto-Electronic Advances*, vol.3, no.9, article no.200012, 2020.

[8] J. W. Ma, "Single-pixel remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol.6, no.2, pp.199–203, 2009.

[9] Z. B. Zhang, X. Y. Wang, G. A. Zheng, *et al.*, "Fast Fourier single-pixel imaging via binary illumination," *Scientific Reports*, vol.7, no.1, article no.12029, 2017.

[10] M. F. Duarte, M. A. Davenport, D. Takhar, *et al.*, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Magazine*, vol.25, no.2, pp.83–91, 2008.

[11] C. B. Li, "An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing," *Master Thesis*, Rice University, Houston, TX, USA, 2010.

[12] M. Lyu, W. Wang, H. Wang, *et al.*, "Deep-learning-based ghost imaging," *Scientific Reports*, vol.7, no.1, article no.17865, 2017.

[13] Y. C. He, G. Wang, G. X. Dong, *et al.*, "Ghost imaging based on deep learning," *Scientific Reports*, vol.8, no.1, article no.6469, 2018.

[14] L. S. Sui, L. W. Zhang, Y. Cheng, *et al.*, "Computational ghost imaging based on the conditional adversarial network," *Optics Communications*, vol.492, article no.126982, 2021.

[15] T. Wolf, L. Debut, 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*, Online, pp.38–45, 2020.

[16] K. Han, Y. H. Wang, H. T. Chen, *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.45, no.1, pp.87–110, 2023.

[17] M. H. Guo, T. X. Xu, J. J. Liu, *et al.*, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol.8, no.3, pp.331–368, 2022.

[18] W. L. Gong and S. S. Han, "A method to improve the visibility of ghost images obtained by thermal light," *Physics Letters A*, vol.374, no.8, pp.1005–1008, 2010.

[19] F. Ferri, D. Magatti, L. A. Lugiato, *et al.*, "Differential ghost imaging," *Physical Review Letters*, vol.104, no.25, article no.253603, 2010.

[20] L. H. Bian, J. L. Suo, Q. H. Dai, *et al.*, "Experimental comparison of single-pixel imaging algorithms," *Journal of the Optical Society of America A*, vol.35, no.1, pp.78–87, 2018.

[21] K. K. Guo, S. W. Jiang, and G. A. Zheng, "Multilayer fluorescence imaging on a single-pixel detector," *Biomedical Optics Express*, vol.7, no.7, pp.2425–2431, 2016.

[22] C. F. Higham, R. Murray-Smith, M. J. Padgett, *et al.*, "Deep learning for real-time single-pixel video," *Scientific Reports*, vol.8, no.1, article no.2369, 2018.

[23] F. Wang, H. Wang, H. C. Wang, *et al.*, "Learning from simulation: An end-to-end deep-learning approach for computational ghost imaging," *Optics Express*, vol.27, no.18, pp.25560–25572, 2019.

[24] H. Wu, R. Z. Wang, G. P. Zhao, *et al.*, "Deep-learning denoising computational ghost imaging," *Optics and Lasers in Engineering*, vol.134, article no.106183, 2020.

[25] I. Hoshi, T. Shimobaba, T. Kakue, *et al.*, "Single-pixel imaging using a recurrent neural network combined with convolutional layers," *Optics Express*, vol.28, no.23, pp.34069–34078, 2020.

[26] X. G. Wang, A. G. Zhu, S. S. Lin, *et al.*, "Learning-based high-quality image recovery from 1D signals obtained by single-pixel imaging," *Optics Communications*, vol.521, article no.128571, 2022.

[27] T. Bian, Y. X. Yi, J. L. Hu, *et al.*, "A residual-based deep learning approach for ghost imaging," *Scientific Reports*, vol.10, no.1, article no.12149, 2020.

[28] T. Bian, Y. M. Dai, J. L. Hu, *et al.*, "Ghost imaging based on asymmetric learning," *Applied Optics*, vol.59, no.30, pp.9548–9552, 2020.

[29] W. Feng, X. Y. Sun, X. H. Li, *et al.*, "High-speed computational ghost imaging based on an auto-encoder network under low sampling rate," *Applied Optics*, vol.60, no.16, pp.4591–4598, 2021.

[30] Y. Fu, T. Zhang, L. Z. Wang, *et al.*, "Coded hyperspectral image reconstruction using deep external and internal learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.44, no.7, pp.3404–3420, 2021.

[31] F. Wang, C. L. Wang, C. J. Deng, *et al.*, "Single-pixel imaging using physics enhanced deep learning," *Photonics Research*, vol.10, no.1, pp.104–110, 2022.

[32] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp.6000–6010, 2017.

[33] A. Gillioz, J. Casas, E. Mugellini, *et al.*, "Overview of the transformer-based models for NLP tasks," in *Proceedings of the 2020 15th Conference on Computer Science and Information Systems*, Sofia, Bulgaria, pp.179–183, 2020.

[34] C. F. R. Chen, Q. F. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp.347–356, 2021.

[35] B. W. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proceedings of the 35th Advances in Neural Information Processing Systems*, Virtual Conference, pp.17864–17875, 2021.

[36] N. Carion, F. Massa, G. Synnaeve, *et al.*, "End-to-end object detection with transformers," in *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp.213–229, 2020.

[37] Y. H. Cai, Z. C. Wang, Z. X. Luo, *et al.*, "Learning delicate local representations for multi-person pose estimation," in *Proceedings of the 16th European Conference on Computer Vision*, Glasgow, UK, pp.455–472, 2020.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint*, arXiv: 2010.11929, 2020.

[39] W. H. Wang, E. Z. Xie, X. Li, *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp.548–558, 2021.

[40] Z. Liu, Y. T. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, QC, Canada, pp.9992–10002, 2021.

[41] S. Wu, J. Tian, and W. Cui, "A novel parameter estimation algorithm for DSSS signals based on compressed sensing," *Chinese Journal of Electronics*, vol.24, no.2, pp.434–438, 2015.

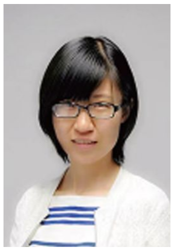
[42] X. S. Wang, Y. H. Cheng, and J. Ji, "Semi-supervised regression algorithm based on optimal combined graph," *Chinese Journal of Electronics*, vol.22, no.4, pp.724–728, 2013.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, Available at: *https://arxiv.org/pdf/1412.6980v9.pdf*, 2015.

[44] S. S. Channappayya, A. C. Bovik, and R. W. Heath, "Rate bounds on SSIM index of quantized images," *IEEE Transactions on Image Processing*, vol.17, no.9, pp.1624–1639, 2008.

[45] K. Zhang, Y. W. Li, W. M. Zuo, *et al.*, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.44, no.10, pp.6360–6376, 2022.

TIAN Ye received the B.S. degree from School of Information Science and Engineering, Lanzhou University, Lanzhou, China, in 2017, the M.S. degree from Peking University, Beijing, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. Her current research interests include deep learning, image processing, and computational imaging. (Email: 3220205110@bit.edu.cn)



FU Ying (corresponding author) received the B.S. degree in electronic engineering from Xidian University in 2009, the M.S. degree in automation from Tsinghua University in 2012, and the Ph.D. degree in information science and technology from the University of Tokyo in 2015. She is currently a Professor at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include physics-based vision, image processing, and computational photography. (Email: fuying@bit.edu.cn)

ZHANG Jun received the B.S., M.S., and Ph.D. degrees in communications and electronic systems from Beihang University, Beijing, China, in 1987, 1991, and 2001, respectively. He was a Professor with Beihang University. He has served as the Dean for the School of Electronic and Information Engineering, and the Vice President and the Secretary for the Party Committee, Beihang University. He is currently a Professor with Beijing Institute of Technology, where he is also the President. His research interests are networked and collaborative air traffic management systems, covering signal processing, integrated and heterogeneous networks, and wireless communications. He is a Member of the Chinese Academy of Engineering. He has won the awards for science and technology in China many times. (Email: buaazhangjun@vip.sina.com)