Incoherent reconstruction-free object recognition with mask-based lensless optics and Transformer

Article *in* Optics Express · October 2021

DOI: 10.1364/OE.443181

4 authors, including:

Xiuxi Pan
Tokyo Institute of Technology
7 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Chen Xiao
Tokyo Institute of Technology
6 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Masahiro Yamaguchi
Tokyo Institute of Technology
353 PUBLICATIONS **4,446** CITATIONS
SEE PROFILE

Research Article | Vol. 29, No. 23 / 8 Nov 2021 / Optics Express 37962
Optics EXPRESS

Incoherent reconstruction-free object recognition with mask-based lensless optics and the Transformer

XIUXI PAN,[1,*] XIAO CHEN,[1] TOMOYA NAKAMURA,[2] AND MASAHIRO YAMAGUCHI[1]

[1]School of Engineering, Tokyo Institute of Technology, 4259-G2-28 Nagatsuta, Midori-ku, Yokohama, Kanagawa 226-8502, Japan
[2]SANKEN, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
*pan.x.aa@m.titech.ac.jp

Abstract: A mask-based lensless camera adopts a thin mask to optically encode the scene and records the encoded pattern on an image sensor. The lensless camera can be thinner, lighter and cheaper than the lensed camera. But additional computation is required to reconstruct an image from the encoded pattern. Considering that the significant application of the lensless camera could be inference, we propose to perform object recognition directly on the encoded pattern. Avoiding image reconstruction not only saves computational resources but also averts errors and artifacts in reconstruction. We theoretically analyze multiplexing property in mask-based lensless optics which maps local information in the scene to overlapping global information in the encoded pattern. To better extract global features, we propose a simplified Transformer-based architecture. This is the first time to study Transformer-based architecture for encoded pattern recognition in mask-based lensless optics. In the optical experiment, the proposed system achieves 91.47% accuracy on the Fashion MNIST and 96.64% ROC AUC on the cats-vs-dogs dataset. The feasibility of physical object recognition is also evaluated.

1. Introduction

Driven by the popularity of Internet of Things (IoT), this decade has seen a surge of demand for smaller, lighter and cheaper cameras that can be applied in extreme scenarios where stringent constraints on size, weight or cost are imposed. Cost down and volume decrease for a lensed camera are restricted by the focusing distance required by refractive lenses. Researchers have been studying mask-based lensless camera which replaces complex lens system by a thin mask [1–12]. Avoiding using lens components, a mask-based lensless camera can extremely reduce the size, weight and cost of optical hardware. However, a mask-based lensless camera moves the imaging burden from optical hardware to computation. It requires additional computation to reconstruct a scene-resembling image from the raw sensor measurements.

With the recent proliferation of artificial intelligence (AI), in many situations, the ultimate goal of visual information acquisition is transferred from human visual appreciation to machine inference. Based on this observation, image acquisition systems that are dedicated to machine inference have been proposed to explore unique features [13–28]. In this work, we propose a dedicated object recognition system with mask-based lensless optical hardware, aiming at relieving imaging burdens in optics without increasing the computation. The mask-based lensless optical hardware consists of a thin mask placed in front of an image sensor. The mask optically encodes the target and produces an encoded pattern on the sensor. Though the encoded pattern is not human-interpretable, it contains visual information of the target that allows object recognition. We argue that image reconstruction is not intrinsically needed in terms of machine inference. Image reconstruction not only uses additional computational resources but also brings errors

#443181 https://doi.org/10.1364/OE.443181
Journal © 2021 Received 13 Sep 2021; revised 4 Oct 2021; accepted 9 Oct 2021; published 29 Oct 2021

Optics EXPRESS

and artifacts. In addition, without resembling the scene, the reconstruction-free strategy can also provide optical-level privacy protection for privacy-sensitive inference tasks such as secure optical sensing [29–31] or de-identified attribute recognition like gender or age estimation. Therefore, we propose to directly perform object recognition on the encoded pattern, bypassing tedious image reconstruction. Figure 1 illustrates the comparison among the proposed system, lensed camera and reconstruction-including mask-based lensless camera considering application in object recognition. The proposed system is ultra-thin, lightweight and low-cost in optical hardware compared with a lensed camera. It is more efficient in both prediction accuracy and computation, compared with reconstruction-including mask-based lensless camera.

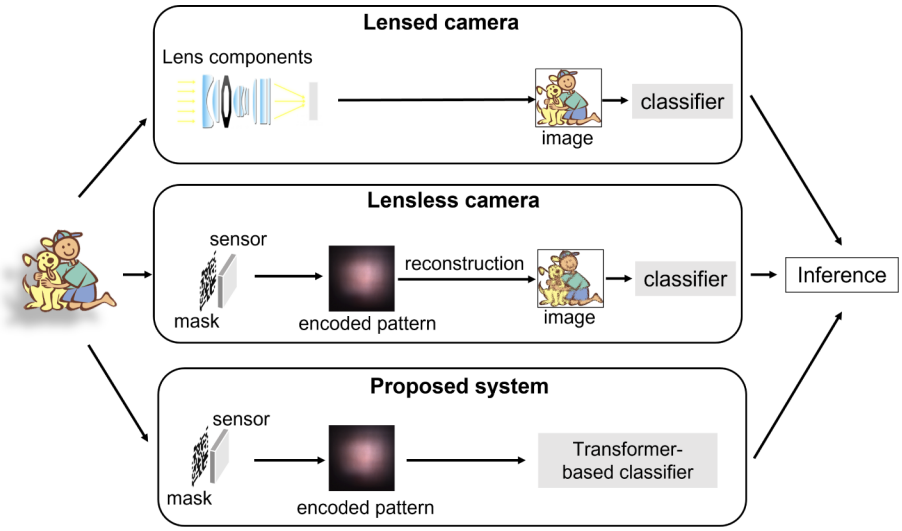

Fig. 1. Comparison among the proposed system, lensed camera and reconstruction-including mask-based lensless camera considering application in object recognition. The proposed system simplifies optical hardware or computation. Avoiding noises and artifacts produced during reconstruction, the proposed system also presents higher prediction accuracy than reconstruction-including mask-based lensless camera.

The lensed camera pursues scene-resembling imaging. Different from that, the mask-based lensless camera maps one point to many pixels. Specifically, through the mask, each point in the scene casts an extensive and specific pattern on the sensor. Herein, local information in the scene is transformed into overlapping global information in the encoded pattern. Consequently, global feature extraction is essential for encoded pattern recognition. The fully convolutional neural network (FCN) [32–36], which is currently dominant machine learning method for computer vision, is not efficient in global feature extraction because it primarily encodes features with stacking small-kernel convolutions. The Transformer-based architecture [37–40] can deal with global features well because it abandons the inductive bias of locality and leverages self-attention mechanism for encoding long-range dependencies. Accordingly, we utilize the Transformer-based architecture rather than convolution-based architecture. Without inductive prior, the power of Transformer relies on a large amount of training data [40]. Capturing millions of encoded patterns with the lensless camera is impractical. To handle this issue, we propose two solutions. Firstly, we apply separated convolutions and axial-attentions [41,42] to computationally simplify the architecture. It makes the training process easier. Secondly, we follow the forward model of mask-based lensless optics, to generate a massive simulated encoded pattern dataset for pretraining. In the end, with large-scale pretraining, the proposed simplified Transformer-based architecture generalizes well in the encoded pattern recognition.

Optics EXPRESS

2. Related work: reconstruction-free recognition

Reconstruction-free recognition is attracting lots of interest and has been applied in a wide range of fields. Speckle patterns produced by coherent laser exposure can provide information for inference tasks with no necessity of image reconstruction [13–20]. Speckle pattern recognition has been used in biomedicine [14–17], agricultural crops investigation [18] and non-line-of-sight recognition [19,20]. Action recognition without image reconstruction in compressive video sensing gains advantages of privacy-preserving [21] and computation reduction [22]. Reconstruction-free recognition has also been introduced to the single-pixel camera for reducing measurement rates or saving computation and sensor pixels [23–27].

Our work considers an incoherent optical system with visible light, which is different from coherent laser illuminated speckle recognition. As we focus on object recognition, the recognition technique presented in our work differs from that used in reconstruction-free action recognition. Compared with the single-pixel camera where a digital mirror device is required, our proposed system is simple in optical hardware and allows single-shot inference.

Lensless inference (LLI) camera [28] pursues incoherent object recognition through a thin mask, which deals with the same problem with this work. LLI camera proposes a data preprocessing approach named "local binary pattern (LBP) map generation" to improve encoded pattern's robustness to disturbances. It is theoretically designed for the cropped and aligned target with a flat background, and doses not consider more complicated scenes. This work proposes to use Transformer-based architecture which is more efficient in global feature reasoning, allowing object recognition in complicated background without cropping and aligning.

3. Mask-based lensless camera: forward model and image reconstruction

Mask-based lensless imaging has been traditionally used in astronomy for x-ray and γ-ray imaging where lenses or mirrors are expensive or infeasible [43,44]. With the recent growing demand for miniature cameras, mask-based lensless imaging has been extended to visible spectrum.

Mask-based lensless imaging, in optical hardware, is simply the axial stack of a thin mask and an image sensor. The mask can be amplitude mask [2,5,7–9,12], phase mask [1,3,4,6,10,11] or any optical encoding element. When the mask is illuminated by a point light source, the sensor acquires a deterministic pattern, called point spread function (PSF). To model the mask-based lensless optics as a linear system, two assumptions are given: (1) the object is a collection of incoherent points with varying intensity, and (2) the PSF is shift-invariant. With these two assumptions, the sensor measurement can be described as the sum of weighted PSFs corresponding to the intensity of the points. Mathematically, the sensor measurement $\mathbf{x}$ is expressed as a linear equation [45]:

$$\mathbf{x} = \Phi\mathbf{o} + \mathbf{e}, \tag{1}$$

where $\mathbf{o}$ is the object, Φ is the measurement matrix constructed by the PSF, and $\mathbf{e}$ is noise. Algorithms are employed to reconstruct the image afterward.

As for image reconstruction methods, there are the iterative optimization [1–6], Moiré-based decoding method [7–9] and learned reconstruction with neural network [10–12]. Although the iterative optimization is computationally demanding, it is the most commonly used because it generally produces better reconstruction quality than others. The iterative optimization method iteratively minimizes a loss function with convex optimization [46–48]. The minimization problem is set as

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmin}} \|\mathbf{x} - \Phi\mathbf{o}\|_{\ell_2}. \tag{2}$$

In the mask-based lensless optics, a sensor pixel measures multiplexed light from widely spread points in the scene. It may often result in an ill-conditioned system. To suppress noise

Optics EXPRESS

amplification, a regularization based on image sparsity prior should be added to the minimization problem

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmin}} \|\mathbf{x} - \Phi\mathbf{o}\|_{\ell_2} + \tau\Psi(\mathbf{o}), \tag{3}$$

where Ψ denotes a regularizer and τ is a parameter tuning the weight of the sparsity prior. We conclude three factors that introduces errors or artifacts in the iterative optimization method. Firstly, the regularization imports data that is not necessarily representative of the original object, especially when the sparsity prior is weak for some scenes. Secondly, the basic assumption of shift-invariance for PSF is broken when there is incident light with high angle. Thirdly, the quality of the captured PSF greatly affects the reconstruction result. The PSF is expected to be captured by illuminating an ideal point light source on the mask. Nonetheless, an ideal point light source can hardly be acquired experimentally.

Moiré-based decoding method accelerates reconstruction, but it is sensitive to noise and calibration precision. Learned reconstruction with neural network learns solutions from the large-scale dataset. This method has potential to fastly reconstruct high-fidelity images. However, for pixel-level image reconstruction, the performance sharply degenerates when the test data deviates from the used training data.

As raw sensor measurements already contain visual information of the target, performing inference on the raw data without image reconstruction is possible. With the above analysis on existing reconstruction algorithms, we think that bypassing reconstruction will not only save computation but also be beneficial in the prediction.

4. Methodology

In Sec.4.1, we firstly illustrate that global feature extraction is critical in encoded pattern recognition due to the multiplexing property in the mask-based lensless optics. With this observation, we propose to utilize the Transformer-based architecture which straightly deals with global context. However, the Transformer requires large-scale datasets for training. To mitigate this issue, we propose the simplified Transformer-based architecture (Sec.4.2) and the method to generate simulated encoded patterns for pretraining (Sec.4.3).

4.1. Global features in the encoded pattern

In the mask-based lensless optics, each point in the scene is mapped to an extensive pattern on the sensor and each sensor pixel measures multiplexed light from widely spread points in the scene. This property is known as multiplexing. Figure 2 illustrates a simulation example of the multiplexing property. The target is a two-dimensional (2D) object and the mask is a binary amplitude mask with a 4×4 array. For simplicity, the 4×4 array of the mask is used as the PSF and the acquired encoded pattern is modeled as the convolution between the object and the PSF. The convolution kernel is large and local information in the object is transformed into global information in the sensor measurement. Accordingly, feature extraction in the scene-resembling image and the encoded pattern should be different. To indicate an object in the scene-resembling image, local features like edges, ridges and corners which focus on a small group of pixels, are essential. While in the encoded pattern, global features which describe pixel arrangement of the entire area tend to be more informative. With this observation, we stress global feature extraction when designing classifier.

Stacking small filters (such as 3×3 or 7×7) in a deep network has been proven efficient in both computation and prediction result [35,36]. Through local area perception, weight sharing and spatial sampling, a typical FCN architecture with stacking small filters initially learns basic local features, and then develop them to high-level abstractions. However, since only small filters are used, global features can only be extracted in deeper layers based on local features extracted in lower layers. As for the encoded pattern, global features are more useful than local features.

Optics EXPRESS

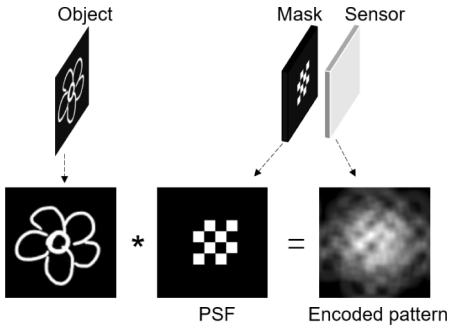

Fig. 2. A simulation example of the forward model in a mask-based lensless optics. Local information in the object is transformed into global information in the encoded pattern.

For better global feature extraction, we propose the use of Transformer-based architecture rather than the FCN. The Transformer-based architecture is free of the inductive bias of locality and considers global context straightly by self-attention mechanism, thus enhancing global feature reasoning.

4.2. *Proposed architecture*

The proposed architecture is mainly comprised of the patchify stem and the Transformer. The patchify stem reshapes the input image to non-overlapping patches, which feed into the Transformer. For computational simplification, large convolution is separated in patchify stem and axial-attentions is used in the Transformer encoder. Figure 3 gives a diagrammatic overview of the proposed architecture.

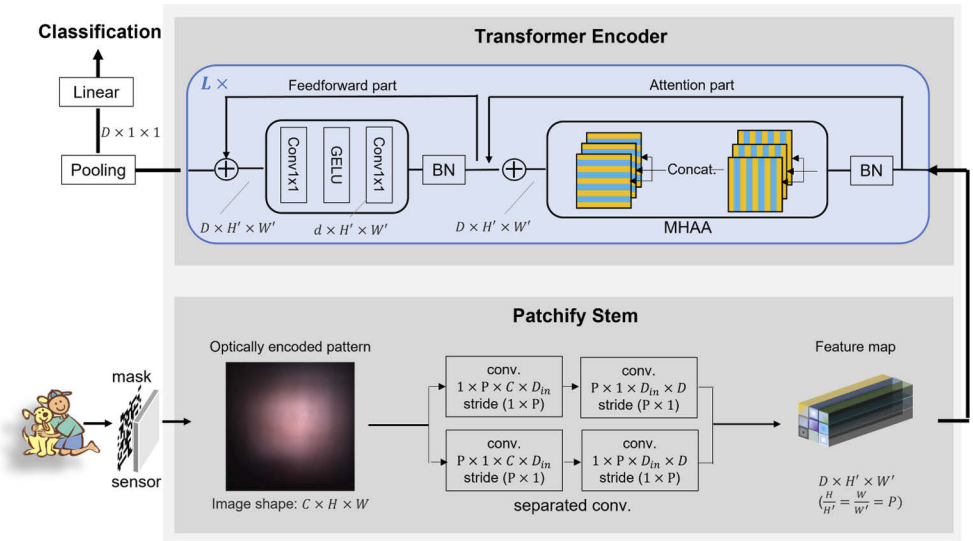

Fig. 3. Diagrammatic overview of proposed architecture. The "BN" refers to batch normalization, "MHAA" refers to multiheaded axial-attentions block and "GELU" refers to Gaussian error linear unit.

Directly applying the Transformer to the input image is impractical considering huge computation occurred. The patchify stem is firstly applied to the input image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$ to get a feature map $\mathbf{x}' \in \mathbb{R}^{D \times H' \times W'}$ with reduced size, where C, D are channel number, H, H' are height and

Optics EXPRESS

W, W' are weight. It is implemented by performing stride-$(P \times P)$, kernel-$(P \times P)$ convolutions on the input image, where $\frac{H}{H'} = \frac{W}{W'} = P$. To considerably reduce the size of feature map, the kernel-$(P \times P)$ is usually large compared with typical small convolution used in FCNs. Large convolution causes the training process volatile [49]. To avoid this issue, the large convolution $conv^{D}_{P\times P\times C}$ with shape of $P \times P \times C$ and number of D is decomposed into separated convolutions $conv_{sep}$. Applying $conv_{sep}$ to the input image $\mathbf{x}$ is expressed as below [50]

$$\begin{aligned}\mathbf{x}' &= conv_{sep}(\mathbf{x}) \\ &= conv^{D}_{P\times 1\times D_{in}}(conv^{D_{in}}_{1\times P\times C}(\mathbf{x})) + conv^{D}_{1\times P\times D_{in}}(conv^{D_{in}}_{P\times 1\times C}(\mathbf{x})),\end{aligned} \tag{4}$$

where D_{in} is set much smaller than D for parameter number reduction. Compared with $conv^{D}_{P\times P\times C}$, $conv_{sep}$ reduces the parameter number from $P \times P \times C \times D$ to $2 \times P \times D_{in} \times (C + D)$.

The proposed Transformer is built on L successive encoders. Each encoder consists of the attention part and the feedforward part. Residual connections are applied to both parts. The attention part includes batch normalization (BN) and multiheaded axial-attentions block (MHAA):

$$a_{l-1} = MHAA(BN(z_{l-1})) + z_{l-1}, \tag{5}$$

and the feedforward part includes BN, stride-(1×1), kernel-(1×1), number-d convolution $conv^{d}_{1\times 1}$, Gaussian error linear unit (GELU), and stride-(1×1), kernel-(1×1), number-$D conv^{D}_{1\times 1}$ in order:

$$z_l = conv^{D}_{1\times 1}(GELU(conv^{d}_{1\times 1}(BN(a_{l-1})))) + a_{l-1}, \tag{6}$$

where $l \in \{1, \ldots, L\}$ is the encoder number, z_{l-1} and z_l are the output of $(l-1)$'th and l'th encoder respectively, a_{l-1} is the output of the attention part with z_{l-1} as input.

In MHAA, n axial-attentions (AA) are computed in parallel and finally concatenated. It is expressed as below considering an input i

$$MHAA(i) = [AA_1(i); AA_2(i); \ldots; AA_n(i)]. \tag{7}$$

Axial-attentions consequently performs self-attention on the feature map width axis and height axis, formally [42]:

$$y_o = \sum_{p\in N_{1\times m(o)}} softmax_p(q_o^T k_p + q_o^T r^q_{p-o} + k_p^T r^k_{p-o})(v_p + r^v_{p-o}). \tag{8}$$

Queries $\mathbf{q} = \mathbf{W_Q i}$, keys $\mathbf{k} = \mathbf{W_K i}$, and values $\mathbf{v} = \mathbf{W_V i}$ are linear projections of the input $\mathbf{i}$. $r^q_{p-o}, r^k_{p-o}, r^v_{p-o}$ are learnable parameters, measuring the compatibility from position p to o in queries, keys and values. The $softmax_p$ signifies a softmax function applied to all possible p positions. $N_{1\times m}(o)$ defines the row or the column where position o exists. y_o is the output at position o. $m = W'$ for the width-wise axial attention, and $m = H'$ for the height-wise axial attention. Different from traditional self-attention which processes the data as a sequence of embeddings, axial-attentions regards the data as graphics. Taking advantage of geometry construction information, axial-attentions reduces the computational complexity by processing pixels in width and height axis separately. Replacing self-attention with axial-attentions reduces the computational complexity from $O(H'^2 W'^2)$ to $O(H'W'm)$.

In the end, the Transformer followed by an average pooling layer and a linear projection:

$$result = linear(pool(\mathbf{z})), \tag{9}$$

where $\mathbf{z} \in \mathbb{R}^{D\times H'\times W'}$ is the output of the Transformer, $\mathbf{z}$ is pooled to a 1D feature with the shape of $D \times 1 \times 1$.

Optics EXPRESS

4.3. Simulated encoded pattern dataset generation

Lacking inductive biases inherent to FCNs, the Transformer do not generalize well when the amount of training data is insufficient [40]. Though this issue is alleviated in the proposed simplified Transformer-based architecture, the amount of needed training data should be still much large compared with training FCNs. For the Transformer, pretraining on large scale datasets before training on the target dataset has been the routine procedure. Pretraining can help to optimally initialize the weights which contributes to higher learning efficiency in the target dataset with relatively small scale. Accordingly, pretraining is also wanted here. However, we find that for encoded pattern recognition, the pretraining on normal image datasets does not work because encoded pattern and normal image differ greatly in the regularity of pixel arrangement, as introduced in Sec.4.1. Encoded pattern datasets for pretraining is demanding. Nonetheless, not like normal images which can be pulled off the internet, there is no available encoded pattern dataset that are ready for use. We propose to generate simulated encoded pattern dataset from available normal image dataset by applying the forward model of mask-based lensless optics.

By assuming that all coming light is able to cast complete pattern on the sensor, Eq. (1) is simplified as

$$\mathbf{x} = \mathbf{o} * \mathbf{a}. \tag{10}$$

The encoded pattern $\mathbf{x}$ is approximated as a convolution between the object $\mathbf{o}$ and the PSF $\mathbf{a}$. To generate simulated encoded pattern, we use normal 2D image as the object, the PSF is captured by illuminating the lensless camera with a point light source. For computational efficiency, we implement Eq. (10) in frequency domain

$$\mathbf{x} = \mathcal{F}^{-1}(\mathcal{F}\mathbf{o}\,\mathcal{F}\mathbf{a}), \tag{11}$$

with $\mathcal{F}$ and $\mathcal{F}^{-1}$ being Fourier transform and inverse Fourier transform. Before performing Fourier transform, zero-padding is applied in both normal image and the PSF to keep them the same size. Zero-padding area is set large enough, in order to satisfy the implicitly assumption of periodic behavior in Fourier method [45].

Though the generated simulated encoded pattern is not identical to the realistic one, the regularity of pixel arrangement is the same. The model pretrained on the simulated encoded pattern dataset can be efficiently transferred into the realistic encoded pattern dataset which is slightly different. It allows the future training being performed on a realistic captured encoded pattern dataset with relatively small scale.

5. Experiment

We conduct two optical experiments, as illustrated in Fig. 4. The first experiment compares methods on standard datasets. The image is displayed one by one on a monitor in front of the camera. Met.1 uses the lensed camera. The result of Met.1 is regarded as the baseline of object recognition. Met.2 is the conventional way to apply lensless camera to objection recognition, which reconstructs image before inference. The result of Met.2 is regarded as the baseline of lensless-camera-based methods. As the FCN currently is the state-of-the-art model for both Fashion MNIST and cats-vs-dogs datasets, we choose ResNet-50 [36] as the classifier in Met.1 and Met.2. Met.3 is the previous method for reconstruction-free encoded pattern recognition (LLI camera) [28]. It uses a data preprocessing approach named "LBP map generation", which allows the FCN classifier to produce better performance in encoded pattern recognition. In Met.3, ResNet-50 is also selected as the classifier. For the proposed Transformer-based architecture, the details are listed in Table 1.

In another experiment, we evaluate the feasibility of lensless methods (Met.2, Met.3 and the proposed method) on physical objects. We built a fruits dataset by collecting 8 fruit classes from

Optics EXPRESS

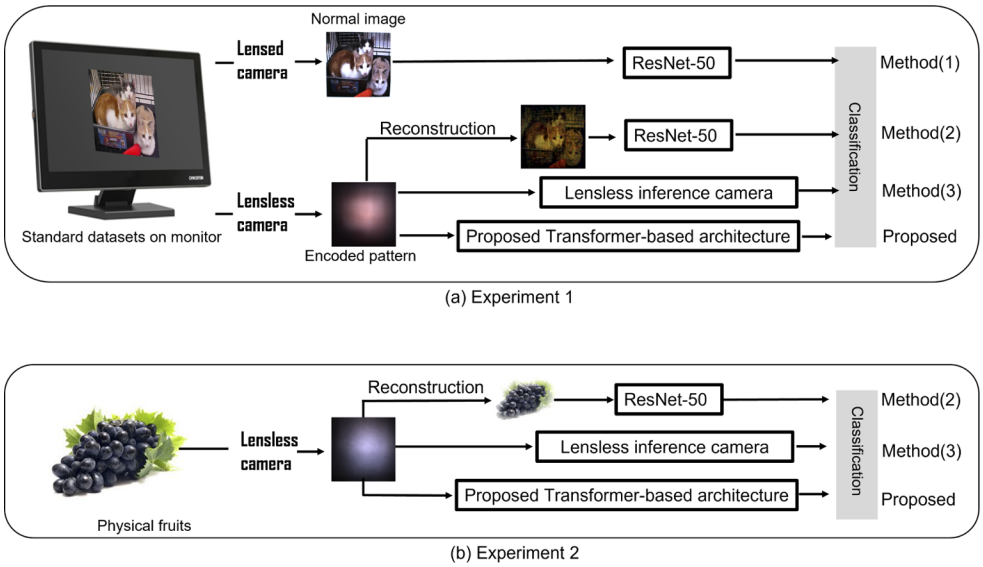

Fig. 4. Diagrammatic overview of two experiments. Exp.1 compares methods on standard datasets. The image is displayed one by one on a monitor in front of the camera. Exp.2 is for evaluating the feasibility of lensless methods on physical objects.

Table 1. Details of the proposed Transformer-based architecture.

	Separated conv.	MHAA	Feedforward
Input Size ($H \times W$)	Patch Size ($P \times P$)	Heads (n)	Inner Depth (d)
224×224	16×16	12	3072
Encoder Num. (L)	Inner Depth (D_{in})		
12	16		
Parameters	Feature Depth (D)		
8.3M	768		

the ILSVRC-2012 ImageNet [51]. We firstly train models as the first experiment. Then physical fruits are used for test.

In Sec.5.1, we elaborate the optical setup of the assembled mask-based lensless hardware. Sec.5.2 introduces experimental details.

5.1. Optical setup of the lensless system

Shown as Fig. 5(b), the mask-based lensless hardware is assembled by a mask and an image sensor, with a separation of 2.5 mm. The used mask is a classic pseudorandom binary amplitude mask without particular optimization, shown as Fig. 5(c). Pseudorandom binary amplitude mask guarantees stable optical signal sampling but does not specifically optimize for any reconstruction or recognition algorithm. It is considered as a fair choice when the experiment aims at comparing algorithms. The amplitude mask is fabricated by chromium deposition in a synthetic-silica plate. It has the optical size of 2.15×2.15 mm and aperture size of 40×40 μ m. The PSF is shown in Fig. 5(d). It is captured when the mask is illuminated by a point LED with the diameter of 1 mm placed 15 cm away. The sensor is a 6.41 megapixels CMOS (Sony IMX178). It has 7.4×5.0 mm optical size, 2.4×2.4 μ m pixel pitch, and records 12 bit color depth. The sensor area is not fully used. Around 1600×1600 sensor pixels are used.

Optics EXPRESS

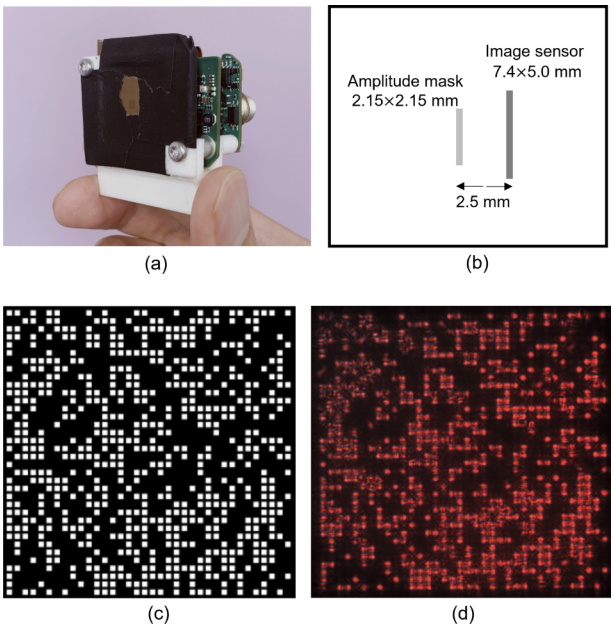

Fig. 5. The mask-based lensless hardware. (a) The photo of assembled mask-based lensless hardware. (b) The schematic diagram. (c) The mask. (d) The PSF, which is captured when the mask is illuminated by a point LED with the diameter of 1 mm placed 15 cm away.

Next, we evaluate the resolution and the degree of diffraction. Considering spatial sampling by the sensor, resolution at the object plane can be given by

$$Resolution = \frac{d_1}{d_2} \times s, \tag{12}$$

where d_1, d_2 and s correspond to target-mask distance, mask-sensor distance and pixel pitch of the sensor respectively. For this lensless hardware, with $d_2 = 2.5$ mm and $s = 2.4$ μm, the object plane resolution is around 0.144 mm when the target is 15 cm away. The degree of diffraction effect can be evaluated by Fresnel number

$$N_F = \frac{a^2}{L\lambda}, \tag{13}$$

where a is the mask aperture size, L is the mask-sensor distance and λ is the incident wavelength. Here, $a = 40$ μ m, $L = 2.5$ mm, and λ is 380~700 nm for visual light. The calculated N_F falls much below 1, which means diffraction effect is considerable comparing with geometrical size. It results in the PSF deviating from the mask, as shown in Fig. 5(d).

5.2. Experimental details

5.2.1. Pretraining

Pretraining is applied for all models in both Exp.1 and Exp.2. ResNet-50 used in Met.1, Met.2 and Met.3 is pretrained on the ILSVRC-2012 ImageNet which has 1000 classes and 1.3 million images in total. The proposed Transformer-based architecture for both Exp.1 and Exp.2 is pretrained on the simulated encoded pattern dataset generated from the ILSVRC-2012 ImageNet.

The simulated encoded pattern is generated on-the-fly during training. The normal image is resized to 224×224 and zero-padded to 448×448. In each epoch, the captured PSF, whose original

Optics EXPRESS

size is 1600×1600, is resized to a size ranging from 150×150 to 224×224 before zero-padded to 448×448. With Eq. (11), a 448×448 encoded pattern is calculated. The center 224×224 area is cropped out for use. In this way, different object-sensor distances are simulated. After encoded pattern is generated, data augmentation techniques including scaling, cropping and horizontal flipping are implemented. An example of simulated encoded pattern is shown in Fig. 6.

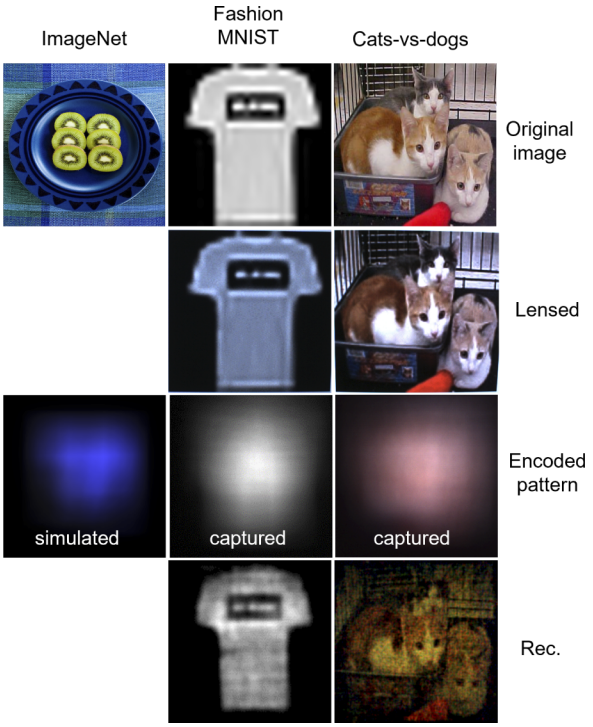

Fig. 6. Examples of used images. In the first row, there are original images from ILSVRC-2012 ImageNet, Fashion MNIST and cats-vs-dogs datasets. In the second row, there are images captured by the lensed camera. In the third row, there are simulated and captured encoded patterns. The last row lists the reconstructed images.

5.2.2. Experiment 1

Exp.1 compares the proposed method with other three methods on standard datasets. The used datasets are Fashion MNIST [52] and cats-vs-dogs dataset [53]. Fashion MNIST dataset consists of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28×28 grayscale image, associated with a label from 10 classes. Cats-vs-dogs dataset consists of 25,000 labeled colorful photos: 12,500 cats and 12,500 dogs. Width and height of each photo range from 100 to 500 pixels. The dataset is split into 80% for training and 20% for testing. Images from the cats-vs-dogs dataset are captured from real environment under complex background without cropping and aligning.

The target is displayed one by one on an LCD monitor with 0.275×0.275 mm pixel pitch, placed around 15 cm away from the camera. All displayed images are resized to 500×500, which occupies around 10×10 cm on the monitor.

In Met.1, the lensed camera uses a 1.3 megapixels sensor (FLIR CM3-U3-13Y3C-S-BD) and a f/1.4 lens. The valid area occupies 600×600 pixels in the captured image. It is cropped out and resized to 224×224 for use.

Optics EXPRESS

For lensless camera, the captured PSF and the encoded pattern have valid area of 1600×1600 pixels. This area is cropped out and resized to 224×224 for use. Met.3 and the proposed method use the encoded pattern directly while Met.2 reconstructs images firstly. In Met.2, the iterative optimization method, which applies alternating direction method of multipliers (ADMM) [48] and total-variation regularization [54] to iteratively solves the optimization problem as Eq. (3), is employed for image reconstruction. Though there are other reconstruction methods, the iterative optimization method is the most mature and most commonly used one. We think it is fair to see it as the representative. The reconstructed image has the size of 224×224, identical to the size of used PSF and encoded pattern. The reconstruction convergence chart, where each point calculates the mean squared error (MSE) between reconstructed images in current iteration and previous iteration, is shown in Fig. 7. Considering convergence situation and computation time, we choose reconstruction with 10 iterations for use. An example of reconstructed image is illustrated in Fig. 6.

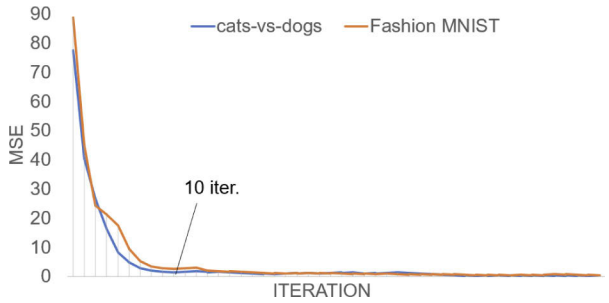

Fig. 7. Reconstruction convergence chart for iterative optimization method. Reconstruction with 10 iterations are chosen for used.

5.2.3. Experiment 2

Exp.2 is designed for evaluating lensless methods' feasibility on physical objects. A fruits dataset is built by collecting 8 fruit classes (apple, banana, grape, kiwi, lemon, orange, peach and watermelon) from the ILSVRC-2012 ImageNet. Each class has around 1,300 images. Models are firstly trained on the fruits data with the same experimental configuration as the Exp.1's. Then test on physical fruits is conducted. We prepare 2 pieces for each fruit kind. As the training images include objects with a wide range of size, there is no specific object-camera distance required in test. To ensure fruits are inside camera view, the object-camera distance ranges from 10 cm to 40 cm based on different fruit sizes.

5.2.4. Training implementation

All trainings are implemented on a machine equipped with an Intel Xeon E5-2698 v4 CPU (2.2 GHz), a NVIDIA TESLA V100 GPU (32 GB), Python 3.6.5, Pytorch 1.7.1. ResNet-50 is trained by Adam optimizer [55] with β_1=0.9, β_2=0.999, a weight decay of 0.1, and a mini-batch size of 64. The proposed Transformer-based architecture is trained by stochastic gradient descent (SGD) optimizer with a learning rate of 0.01, momentum of 0.9 and mini-batch size of 64. The comparison of training processes is illustrated in Table 2 where used epochs for convergence and runtime for one epoch are listed.

Optics EXPRESS

Table 2. Training process comparison.

Task	Method	Epochs	Time/Epoch
Clothing classification	Met.1 (Lensed)	15	1 min
	Met.2 (Rec.)	15	1 min
	Met.3 (LLI [28])	25	1 min
	Proposed	20	4 min 30 s
Cats-vs-dogs classification	Met.1	15	1 min
	Met.2	15	1 min
	Met.3	45	1 min
	Proposed	15	4 min 30 s

6. Result

The result of Exp.1 is listed in Table 3. We compare both accuracy/ ROC AUC (area under the receiver operating characteristic curve) and runtime. All computations, including reconstruction and inference, are timed on an Intel Xeon E5-2698 v4 CPU (2.2 GHz). The proposed method presents higher prediction accuracy while less runtime than reconstruction-including lensless imaging (Met.2). It verifies that image reconstruction in not intrinsically needed in terms of inference. Bypassing reconstruction not only saves computation but also benefits recognition performance. Met.3 is theoretically designed for cropped and aligned target without complex background. It hereby works well for Fashion MNIST where the target is aligned on a black background without noise, while fails on cats-vs-dogs dataset where the target is in complex real-world scene. The result of the proposed method is marginally worse than that of lensed-camera-used method (Met.1). Theoretically, encoded pattern recognition has potential to have the same performance as normal image recognition, because the mask-based lensless optics and lensed camera record the same visual information in different encoding ways. The performance of the proposed method is expected to be closer to the Met.1's if the optical hardware or classifier is further optimized. For example, suppressing diffraction or optimize mask design can result in better optical signal sampling, thereby enhancing recognition. If computational resources permits, smaller patch size and deeper network can also contribute to higher predictive accuracy.

Table 3. Result of Exp.1. Total consumed time for Met.2 is the sum of reconstruction and inference time.

		Acc./AUC (%)	Time (s) for 10 images		
			Rec.	Infer.	Total
Clothing classification	Met.1 (Lensed)	91.50 / -	-	0.99	0.99
	Met.2 (Rec.)	91.12 / -	1.46	0.99	2.45
	Met.3(LLI [28])	90.01 / -	-	0.99	0.99
	Proposed	91.47 / -	-	1.00	1.00
cats-vs-dogs classification	Met.1	97.00 / 97.79	-	1.04	1.04
	Met.2	79.85 / 86.76	4.38	1.04	5.42
	Met.3	74.02 / 82.10	-	1.04	1.04
	Proposed	94.26 / 96.64	-	1.04	1.04

Shooting scenes, captured encoded patterns and reconstructed images of Exp.2 are shown in Fig. 8. The result is illustrated in Table 4, where top three class probability predictions for each item are listed. The proposed method achieves 13/16 accuracy. The confidence is not high for some correct predictions due to domain difference between training data (2D images displayed on

Optics EXPRESS

the monitor) and test data (physical 3D objects). The result validates that the proposed method is able to generalize to physical objects. But for robust prediction in practical application, training with physical objects is needed. For both Met.2 and Met.3, only 3 out of 16 predictions are correct. Met.3 has poor performance because it is theoretically inapplicable for the non-aligned scene as introduced before. For Met.2, the result is much worse than the result in Exp.1. Besides the domain difference, the main reason is that the image reconstructed by the optimization method

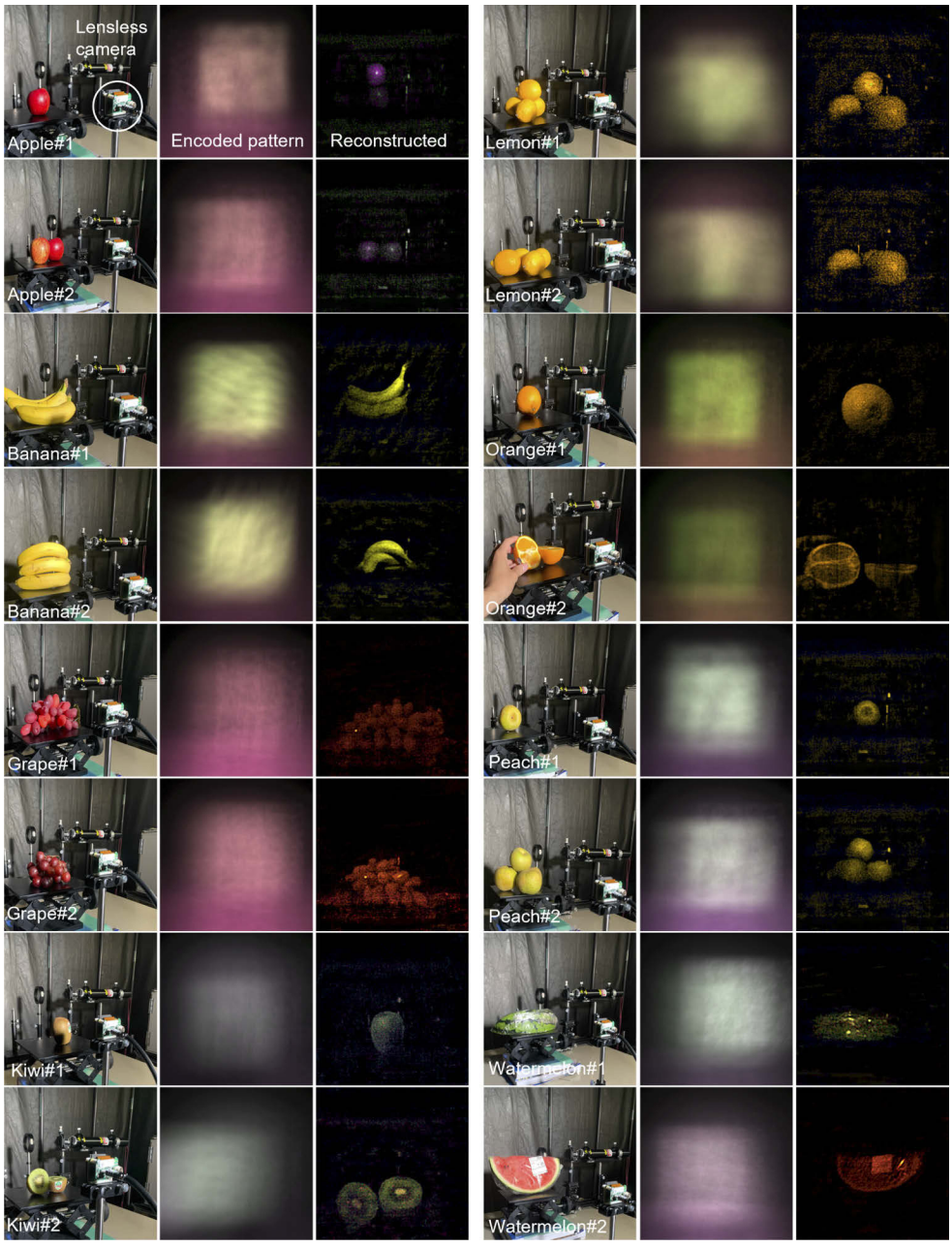

Fig. 8. Shooting scenes, captured encoded patterns and reconstructed images of Exp.2.

Optics EXPRESS

have much lower quality in real world scene. In real world scene where the light conditions is complex, the basic assumption of shift-invariance for PSF can hardly be met as high-angle incident light is inevitable if the position of the target is not carefully selected.

Table 4. Result of Exp.2. Top three class probability predictions for each item are listed. Correct predictions are shown in bold.

Item	Met.2	Met.3	Proposed
Apple 1	Apple: 0.37 Banana: 0.36 Orange :0.12	Apple: 0.83 Others<0.1	Apple: 0.86 Orange: 0.11 Others<0.1
Apple 2	Banana: 0.39 Apple: 0.32 Orange: 0.13	Apple: 0.62 Orange: 0.1 Others<0.1	Apple: 0.63 Orange: 0.19 Watermelon: 0.10
Banana 1	Banana: 0.34 Apple: 0.32 Others<0.1	All<0.1	All<0.1
Banana 2	Watermelon: 0.34 Banana: 0.16 Apple: 0.16	All<0.1	All<0.1
Grape 1	Banana :0.42 Apple: 0.33 Orange: 0.10	Apple: 0.30 Grape: 0.13 Orange: 0.13	Grape: 0.53 Apple: 0.14 Orange: 0.13
Grape 2	Banana: 0.36 Apple: 0.32 Orange: 0.10	Apple: 0.30 Grape: 0.14 Kiwi: 0.13	Grape: 0.29 Apple: 0.26 Watermelon: 0.16
Kiwi 1	Apple: 0.32 Banana: 0.32 Orange:0.15	Watermelon: 0.24 Apple: 0.21 Grape: 0.10	kiwi: 0.44 Watermelon: 0.27 Apple: 0.11
Kiwi 2	Banana: 0.40 Apple: 0.35 Orange: 0.11	Watermelon: 0.34 Apple: 0.32 Kiwi: 0.14	Kiwi: 0.43 Apple: 0.36 Watermelon: 0.12
Lemon 1	Banana: 0.39 Orange: 0.21 Apple: 0.16	Orange: 0.71 Apple: 0.19 Others<0.1	Lemon: 0.55 Orange: 0.27 Peach: 0.10
Lemon 2	Kiwi: 0.37 Banana: 0.24 Apple: 0.19	All<0.1	All<0.1
Orange 1	Orange: 0.45 Apple: 0.25 Banana: 0.16	Orange: 0.50 Others <0.1	Orange: 0.37 Apple: 0.16 Peach: 0.01
Orange 2	Apple: 0.40 Banana: 0.29 Orange: 0.13	All <0.1	Orange: 0.35 Apple: 0.17 Peach: 0.12
Peach 1	Apple: 0.41 Banana: 0.25 Orange: 0.16	Orange: 0.47 Apple: 0.24 Others<0.1	Peach: 0.45 Orange: 0.24 Apple: 0.15
Peach 2	Apple: 0.34 Banana: 0.35 Orange: 0.15	Apple: 0.47 Orange: 0.27 Peach: 0.26	Peach: 0.43 Apple: 0.27 Orange: 0.21
Watermelon 1	Banana: 0.42 Apple: 0.28 Orange: 0.14	Apple: 0.28 Watermelon: 0.17 Banana: 0.14	Watermelon: 0.37 Apple: 0.30 Banana: 0.18
Watermelon 2	Banana: 0.40 Apple: 0.31 Orange: 0.12	Apple: 0.17 Watermelon: 0.15 Grape: 0.11	Watermelon: 0.40 Grape: 0.15 Apple: 0.14

Optics EXPRESS

7. Conclusion

To better serve AI and IoT applications, we propose a lensless, reconstruction-free object recognition system. The proposed system is free of the lens, simply adopts a thin mask to optically encode the target and produce an encoded pattern on the image sensor. Since the mask and the sensor can be fabricated integrally during semiconductor processes, the mask-based lensless camera has the potential to be extremely thin, lightweight and cheap. Different from other studies in the lensless imaging field which focus on developing better reconstruction techniques to recover the image from the encoded pattern, we focus on inference directly on the encoded pattern. In terms of inference, bypassing image reconstruction not only saves computational resources but also averts risk of adding errors and artifacts during reconstruction.

We analyze the multiplexing property in mask-based lensless optics and illustrate that global features are essential for understanding encoded pattern. Based on this fact, a Transformer-based architecture is proposed for encoded pattern recognition. Simulated encoded pattern generation method and architecture simplification techniques are also proposed to address the training data shortage problem.

The optical experiments illustrate that the proposed method is close to the lensed-camera-used method, and outperforms reconstruction-including lensless camera method in both predictive accuracy and computation speed. The feasibility of the proposed method on physical object is also verified with a fruits recognition test.

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data and code underlying the results presented in this paper are available in Ref. [56].

References

1. D. G. Stork and P. R. Gill, "Optical, Mathematical, and Computational Foundations of Lensless Ultra-Miniature Diffractive Imagers and Sensors," International Journal on Advances in Systems and Measurements **7**(3-4), 201–208 (2014).
2. M. J. DeWeert and B. P. Farm, "Lensless coded aperture imaging with separable doubly Toeplitz masks," Opt. Eng. **9109**, 91090Q (2014).
3. S. K. Sahoo, D. Tang, and C. Dang, "Single-shot multispectral imaging with a monochromatic camera," Optica **4**(10), 1209–1213 (2017).
4. N. Antipa, G. Kuo, R. Heckel, B. Mildenhall, E. Bostan, R. Ng, and L. Waller, "Diffusercam: lensless single-exposure 3d imaging," Optica **5**(1), 1–9 (2018).
5. M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "Flatcam: Thin, lensless cameras using coded aperture and computation," IEEE Trans. Comput. Imaging **3**(3), 384–397 (2017).
6. V. Boominathan, J. K. Adams, J. T. Robinson, and A. Veeraraghavan, "Phlatcam: Designed phase-mask based thin lensless camera," IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1618–1629 (2020).
7. T. Shimano, Y. Nakamura, K. Tajima, M. Sao, and T. Hoshizawa, "Lensless light-field imaging with Fresnel zone aperture: quasi-coherent coding," Appl. Opt. **57**(11), 2841–2850 (2018).
8. T. Nakamura, T. Watanabe, S. Igarashi, X. Chen, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, "Superresolved image reconstruction in fza lensless camera by color-channel synthesis," Opt. Express **28**(26), 39137–39155 (2020).
9. X. Chen, T. Nakamura, X. Pan, K. Tajima, K. Yamaguchi, T. Shimano, and M. Yamaguchi, "Resolution improvement in fza lens-less camera by synthesizing images captured with different mask-sensor distances," in *2021 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2021), pp. 2808–2812.
10. Y. Li, Y. Xue, and L. Tian, "Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media," Optica **5**(10), 1181–1190 (2018).
11. K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," Opt. Express **27**(20), 28075–28090 (2019).
12. S. S. Khan, V. Adarsh, V. Boominathan, J. Tan, A. Veeraraghavan, and K. Mitra, "Towards photorealistic reconstruction of highly multiplexed lensless images," in *Proceedings of the IEEE International Conference on Computer Vision*, (2019), pp. 7860–7869.
13. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**(6406), 1004–1008 (2018).
14. Z. Zalevsky, Y. Beiderman, I. Margalit, S. Gingold, M. Teicher, V. Mico, and J. Garcia, "Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern," Opt. Express **17**(24), 21566–21580 (2009).

Optics EXPRESS

15. B. Javidi, S. Rawat, S. Komatsu, and A. Markman, "Cell identification using single beam lensless imaging with pseudo-random phase encoding," Opt. Lett. **41**(15), 3663–3666 (2016).
16. B. Javidi, A. Markman, and S. Rawat, "Automatic multicell identification using a compact lensless single and double random phase encoding system," Appl. Opt. **57**(7), B190–B196 (2018).
17. T. O'Connor, C. Hawxhurst, L. M. Shor, and B. Javidi, "Red blood cell classification in lensless single random phase encoding using convolutional neural networks," Opt. Express **28**(22), 33504–33515 (2020).
18. A. Zdunek, A. Adamiak, P. M. Pieczywek, and A. Kurenda, "The biospeckle method for the investigation of agricultural crops: A review," Opt. Lasers Eng. **52**, 276–285 (2014).
19. X. Lei, L. He, Y. Tan, K. X. Wang, X. Wang, Y. Du, S. Fan, and Z. Yu, "Direct object recognition without line-of-sight using optical coherence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), pp. 11737–11746.
20. M. Isogawa, Y. Yuan, M. O'Toole, and K. M. Kitani, "Optical non-line-of-sight physics-based 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), pp. 7013–7022.
21. Z. W. Wang, V. Vineet, F. Pittaluga, S. N. Sinha, O. Cossairt, and S. Bing Kang, "Privacy-preserving action recognition using coded aperture videos," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW), (2019).
22. T. Okawara, M. Yoshida, H. Nagahara, and Y. Yagi, "Action recognition from a single coded image," in *2020 IEEE International Conference on Computational Photography (ICCP)*, (IEEE, 2020), pp. 1–11.
23. M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Computational Imaging V*, vol. 6498 (International Society for Optics and Photonics, 2007), p. 64980H.
24. S. Lohit, K. Kulkarni, P. Turaga, J. Wang, and A. C. Sankaranarayanan, "Reconstruction-free inference on compressive measurements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2015), pp. 16–24.
25. K. Kulkarni and P. Turaga, "Reconstruction-free action inference from compressive imagers," IEEE Trans. Pattern Anal. Mach. Intell. **38**(4), 772–784 (2016).
26. S. Jiao, J. Feng, Y. Gao, T. Lei, Z. Xie, and X. Yuan, "Optical machine learning with incoherent light and a single-pixel detector," Opt. Lett. **44**(21), 5186–5189 (2019).
27. Z. Zhang, X. Li, S. Zheng, M. Yao, G. Zheng, and J. Zhong, "Image-free classification of fast-moving objects using learned structured illumination and single-pixel detection," Opt. Express **28**(9), 13269–13278 (2020).
28. X. Pan, T. Nakamura, X. Chen, and M. Yamaguchi, "Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation," Opt. Express **29**(7), 9758–9771 (2021).
29. B. Javidi and J. L. Horner, "Optical pattern recognition for validation and security verification," Opt. Eng. **33**(6), 1752–1756 (1994).
30. P. Refregier and B. Javidi, "Optical image encryption based on input plane and fourier plane random encoding," Opt. Lett. **20**(7), 767–769 (1995).
31. B. Javidi, A. Carnicer, M. Yamaguchi, T. Nomura, E. Pérez-Cabré, M. S. Millán, N. K. Nishchal, R. Torroba, J. F. Barrera, W. He, X. Peng, A. Stern, Y. Rivenson, A. Alfalou, C. Brosseau, C. Guo, J. T. Sheridan, G. Situ, M. Naruse, T. Matsumoto, I. Juvells, E. Tajahuerce, J. Lancis, W. Chen, X. Chen, P. W. H. Pinkse, A. P. Mosk, and A. Markman, "Roadmap on optical security," J. Opt. **18**(8), 083001 (2016).
32. K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," Neural networks **1**(2), 119–130 (1988).
33. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural computation **1**(4), 541–551 (1989).
34. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems* 25, 1097–1105 (2012).
35. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556 (2014).
36. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 770–778.
37. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, (2017), pp. 5998–6008.
38. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805 (2018).
39. X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 7794–7803.
40. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929 (2020).
41. J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, "Axial attention in multidimensional transformers," arXiv preprint arXiv:1912.12180 (2019).

Optics EXPRESS

42. H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*, (Springer, 2020), pp. 108–126.
43. R. Dicke, "Scatter-hole cameras for x-rays and gamma rays," Astrophys. J. **153**, L101 (1968).
44. E. E. Fenimore and T. M. Cannon, "Coded aperture imaging with uniformly redundant arrays," Appl. Opt. **17**(3), 337–347 (1978).
45. J. W. Goodman, *Introduction to Fourier optics* (Roberts and Company Publishers, 2005).
46. J. M. Bioucas-Dias and M. A. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," IEEE Transactions on Image Processing **16**(12), 2992–3004 (2007).
47. A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," IEEE Transactions on Image Processing **18**(11), 2419–2434 (2009).
48. S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (Now Publishers Inc, 2011).
49. K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "Vitgan: Training gans with vision transformers," arXiv preprint arXiv:2107.04589 (2021).
50. C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), pp. 4353–4361.
51. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, (Ieee, 2009), pp. 248–255.
52. H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747 (2017).
53. O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*, (IEEE, 2012), pp. 3498–3505.
54. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," Phys. D **60**(1-4), 259–268 (1992).
55. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).
56. X. Pan, "Lensless inference transformer repository," Github (2021) [retrieved 2021-10-01], https://github.com/BobPXX/LLI_Transformer.