東京工業大学
Tokyo Institute of Technology

Next-generation Imaging and Sensing: Lensless Optics Equipped with Deep Learning

Xiuxi Pan

Department of Information and Communications Engineering

School of Engineering

A thesis presented for the degree of

Doctor of Engineering

Advisor: Professor Masahiro Yamaguchi

June 2022

Acknowledgements

This thesis is a culmination of work that could not have been accomplished without supports from several people. First, I would like to express my gratitude to my advisor, Prof. Masahiro Yamaguchi, for continuous support and guidance throughout this journey. He is always patient and values my opinions. I have enjoyed sharing, discussing and working with him. I admire his vision, rigorous academic attitude and clear logical analysis; he will continue to inspire and motivate me in my future career.

I also appreciate Prof. Tomoya Nakamura and Prof. Saori Takeyama. I value their guidance and the time we spent together on shaping research ideas and ruminating over technical details. Their mentorship and friendship were insightful and enjoyable. I was lucky to have received guidance from several others as well, including but not limited to my second supervisor Prof. Konstantinos Slavakis, my mentor Prof. Nobuhiro Hayashi at WISE-SSS, my manager Prof. Rui Ishiyama at NEC Data Science Laboratory (internship, Oct 2021 - Dec 2021), and my manager and colleagues Dr. Yoshinori Konishi, Mr. Hiroki Sakuma and Mr. Kenichi Nakazato at SenseTime Japan Ltd. (internship, Mar 2022 - Apr 2022).

My amazing peers, Mr. Xiao Chen, Mr. Sanchez Alexis and others at Yamaguchi Laboratory shared experimental experience with me and provided me with valuable discussion. I appreciate their different perspectives on my work, which improve the research quality.

I sincerely thank the support from Tokyo-Tech Pioneering Doctoral Research Program and Super Smart Society Promotion Consortium.

I would like to express my sincere gratitude to the thesis committee, Prof. Itsuo Kumazawa, Prof. Yoshihiro Watanabe, Prof. Jaehoon Yu and Prof. Konstantinos

Slavakis, for their precious time and intellectual contributions to improving this thesis.

A special gratitude must be expressed to retired Prof. Shinichi Komatsu of Waseda University, who opened the door for me to study in Japan when I was working in a fabrication plant in China. My story would not come here without this turning point.

Finally, my deepest gratitude goes towards the constant, selfless love and support from my family. I would like to thank my parents, Guozhen and Meibin, for all their self-sacrifice on my behalf. My wife Dr. Dongwan Yan is cute, smart and thoughtful, making things easily. Our energetic daughter Qianmo also joined in the first year of my Ph.D. period. My family empowers me to conquer challenges including this thesis.

Abstract

A mask-based lensless camera optically encodes scenes with a thin mask and reconstructs the image with post-capture computations afterward. Free of lens, the lensless camera is naturally small, lightweight and low-cost in optical hardware. However, at present the imaging quality of the lensless camera is unsatisfactory mainly due to limitations in existing image reconstruction methods. The imprecise modeling of the lensless system depreciates the performance of the conventional model-based reconstruction methods. Data-driven deep learning could be an alternative solution because it relies more on data rather than the modeling of the target problem. Nonetheless, the overwhelming majority of deep learning algorithms have been developed for the normal scene-resembling image. They are inadequate or impractical for the optically encoded pattern produced by the lensless camera. This thesis works on dedicated deep learning algorithms for the lensless camera considering the unique multiplexing property in lensless optics. With the proposed deep learning algorithms, the lensless camera has considerable improved imaging quality and additionally unlocks the reconstruction-free recognition functionality. A novel learned mask pattern optimization method is also proposed. This thesis advances the lensless camera to be lite-yet-mighty by equipping the proposed deep learning algorithms.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

LIT	Lensless Imaging Transformer
LRT	Lensless Recognition Transformer
PSF	Point Spread Function
DNN	Deep Neural Network
FCN	Fully Convolutional Neural Network
BN	Batch Normalization
MHAA	Multiheaded Axial-attentions
GELU	Gaussian Error Linear Unit
ADMM	Alternating Direction Method of Multipliers
MSE	Mean Squared Error
PSNR	Peak Signal to Noise Ratio
SSIM	Structural Similarity Index Measure
LBPMG	Local Binary Pattern Map Generation
LBPH	Local Binary Pattern Map Histogram
FZA	Fresnel Zone Apertures

1 | Introduction

1.1 Motivation and Objectives

The lensed camera has been the dominant imaging solution since its invention in the mid-nineteenth century. A lensed camera aims at achieving high-quality, bright and aberration-free imaging by designing a lens system to focus light from the scene onto a photosensitive surface. This decade has seen a rapidly growing demand for smaller, lighter and cheaper cameras that are feasible in extreme scenarios, such as Internet of Things (IoT) devices, where stringent constraints on size, weight and cost are imposed. The optical hardware of the lensed camera is dominated by the lens system and the focusing distance required by refractive lenses. Based on this fact, recently researchers have been studying mask-based lensless cameras for incoherent visible light imaging which simplifies optical hardware by replacing the lens system with a thin mask placed closely in front of the image sensor [1, 2, 3, 4, 5, 6, 7, 8, 9]. Eschewing lens, light through the mask casts an "unknown" smudge, rather than a focused image, on the sensor. This smudge, however, contains sufficient information for reconstructing the image as if a lens system were in place if a well-designed post-capture computational algorithm is applied.

Being miniature and low-cost are obvious advantages; however, by no means all of the lensless optics. Advantages of the mask-based lensless camera are summarized as follows:

1) **Being miniature and low-cost.** Dispensing with lens greatly reduces the size, weight and cost of optical hardware. The mask used is typically an amplitude mask or phase mask. Such masks can be fabricated integrately

as the top layer of the image sensor in semiconductor fabrication. It not only further miniaturizes optical hardware but also avoids additional cost of assembling and calibrating the optical system during production.

2) **Wavelength scaling.** The lensless camera can be theoretically extended to invisible light imaging, such as X-ray, γ-ray, mm-wave, terahertz and thermal wavebands imaging, which is impractical or even impossible for lensed cameras [10, 11].

3) **Single-shot three-dimensional (3D) imaging.** Due to the fact that the point spread functions of the mask-based lensless camera scale in size based on the depth of the point source, the lensless measurements can encode depth information in a certain depth range. With compressed sensing [12, 13, 14], reconstructing more voxels than pixels captured is possible, provided that the 3D sample is sparse in some domain. This theoretically enables one-shot 3D imaging [3] and post-capture refocusing [6].

Considering these advantages, the mask-based lensless optics is a promising alternative form of imaging solution. Nevertheless, the lensless imaging technology is still at its formative stage. This thesis focuses on advancing the development progress of this technology by bringing in deep learning. In particular, three main contributions are achieved in this thesis by equipping mask-based lensless optics with deep learning:

1) **Improved image reconstruction.** Imaging quality improvement is always pursued in imaging technologies including lensless imaging. This thesis studies the physical properties of the mask-based lensless optics and designs the dedicated Transformer-based neural network to achieve improved image reconstruction. The proposed network produces better image quality than existing methods while the computational speed is sufficiently high to

enable real-time shots. A part of the related work was published in [15, 16]. The relevant code is available in [17].

2) **Unlocking reconstruction-free recognition.** This thesis unlocks the functionality of reconstruction-free recognition for the mask-based lensless optics. Today, a large proportion of cameras no longer photograph, instead capture the visual information for computer vision tasks such as object recognition. With this observation, serving for computer vision could be a significant application of the lensless camera. In this thesis, performing object recognition directly on the raw sensor measurements without image reconstruction is proposed for the first time. Avoiding image reconstruction not only saves computational resources, but also averts errors and artifacts caused by reconstruction. This pathbreaking work extends the lensless camera from imaging to sensing. A part of related work was published in [18, 19]. The relevant code is available in [20].

3) **Learned mask pattern optimization.** Mask pattern design is another important subject in lensless camera. A novel learned mask pattern optimization method is proposed. Since the sensor measurement can be approximated as the convolution between the scene and the mask in mask-based lensless optics, the mask array becomes a convolutional layer that can be collectively optimized with the weights of the subsequent neural network.

An additional contribution is addressing non-convex losses issue [21] in Transformer-based neural network training. This thesis proposes to simplify Transformer-based neural network by using axial-attentions (Chapter 3.3) and separated convolutions (Chapter 4.3), and to prepare simulated pretraining data (Chapter 4.3).

1.2 Thesis Outline

The remainder of this thesis is organized as follows.

Chapter 2 — introduces the background of mask-based lensless optics and answers why deep learning is the preferred post-capture computational algorithm for mask-based lensless optics.

Chapter 3 — addresses image reconstruction improvement, which is the most essential subject of the lensless camera. This chapter reviews the existing image reconstruction algorithms and proposes the dedicated Transformer-based neural network, dubbed as the "Lensless Imaging Transformer (LIT)".

Chapter 4 — introduces reconstruction-free recognition, presenting a new functionality for the lensless camera. This chapter first explains the needs and benefits of reconstruction-free recognition as well as difficulties of achieving it. Then, solutions with data preprocessing method and end-to-end neural network design are proposed, respectively. The data preprocessing method is the "Local Binary Pattern Map Generation" developed to work with fully convolutional neural networks. The proposed end-to-end neural network is a Transformer-based neural network, dubbed as the "Lensless Recognition Transformer (LRT)".

Chapter 5 — introduces a unique mask pattern optimization method, which optimizes the aperture distribution through learning by integrating the mask as a convolutional layer ahead of the subsequent neural network.

Chapter 6 — explores potential applications of the lensless camera.

Chapter 7 — provides conclusion and discussion.

2 | Lensless Optics and Deep Learning

2.1 Mask-based Lensless Optics

A mask-based lensless optical system is simply an axial stack of a thin mask and an image sensor, shown as Figure 2.1. The mask can be an amplitude mask, phase mask or any optical encoding element. The invention of the mask-based lensless camera is inspired by the pinhole camera. Since the amount of light researching the senor is restricted by the pinhole, the pinhole camera is inefficient. The mask in the mask-based lensless camera can be seen as an array of pinholes, which improves the light throughput. The sensor measurements become a superposition of images formed by each pinhole, and a post-capture computational algorithm is required to reorganize the measurements. The mask-based lensless imaging has been traditionally used in astronomy for X-ray and γ-ray imaging where lens or mirror are impractical or infeasible [10, 11]. Until recently, mask-based lensless imaging has not been extended to the incoherent visible spectrum [1, 2, 3, 4, 5, 6, 7, 8, 9].

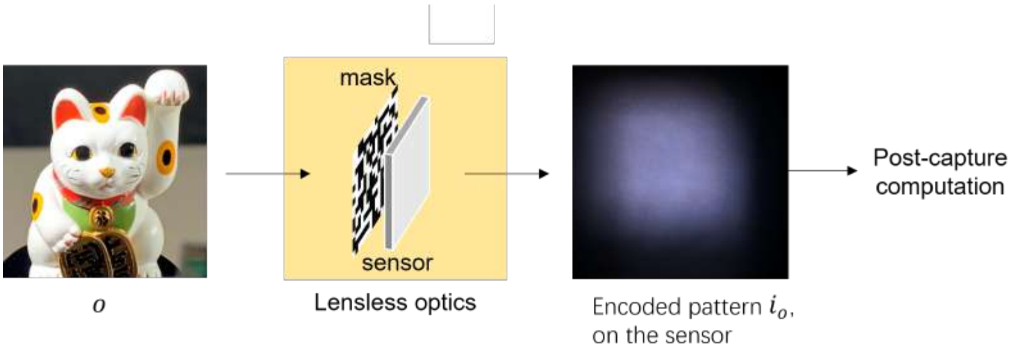


Figure 2.1: Mask-based lensless optics

When the mask is illuminated by a point light source, the sensor acquires a deterministic pattern, called the point spread function (PSF), as illustrated in Figure 2.2. Assuming that the object is a collection of incoherent points with varying intensity, the sensor measurement can be described as the sum of weighted PSFs corresponding to the intensity of the points. Mathematically, the sensor measurement $\mathbf{i_o}$ is expressed as [22]:

$$\mathbf{i_o} = \Phi\mathbf{o} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{i_o}$ is the sensor measurement, $\mathbf{o}$ is the object, Φ is the measurement matrix constructed by the PSF, and $\mathbf{e}$ is noise. If the PSF is shift-invariant, Φ is a block Toeplitz matrix and Eq. 2.1 is a linear function. The sensor measurement $\mathbf{i_o}$ is called the "encoded pattern" because the object is optically encoded by the mask and casts the pattern on the sensor. The encoded pattern is "meaningless" to human eyes; thus post-capture computation is required for image reconstruction or analysis.

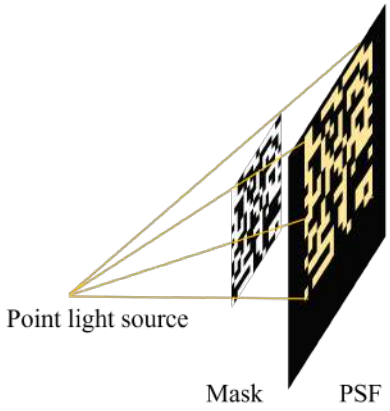


Figure 2.2: PSF of the mask-based lensless optics

Highly efficiently solving or analyzing target object $\mathbf{o}$ from Eq. 2.1 is challenging because the modeling of the system is imprecise. First, Eq. 2.1 is linear only when the PSF is shift-invariant. Shift-invariance means the PSF changes only in location, not in functional form, if the point light moves. In practice, shift-invariance is rarely satisfied over the entire object field if the object is not sufficiently small. Second, an ideal PSF can hardly be acquired. An ideal PSF is measured when the mask is illuminated by an ideal point light source which does not yet exist in practice.

2.2 Equipping Lensless Optics with Deep Learning

Deep Learning is a subfield of machine learning concerned with the algorithms inspired by the structure and function of the brain called artificial neural networks. In traditional machine learning, feature extraction is handled manually, while in deep learning feature extraction happens automatically during the learning process. Compared with traditional machine learning which should be manually designed towards specific cases, deep learning is more generalized. More importantly, pure data-driven deep learning relies more on data than on modeling; thus, it could work even when the modeling of the target problem is unclear or too complicated.

As stated in Section 2.1, the modeling of the mask-based lensless system (Eq. 2.1) is imprecise because the assumption of shift-invariant PSF does not always hold and the ideal PSF cannot be acquired in practice. With imprecise modeling, it becomes hard to manually design highly efficient post-capture computational algorithms. Considering its specialty in dealing with unclear modeling, deep learning is a suitable choice of computational algorithm and could reinvigorate advancements in the mask-based lensless camera.

However, most existing deep learning algorithms have been designed to process normal scene-resembling images. They have limitations in dealing with encoded patterns whose regularity of pixel arrangement is quiet different from the normal image. This thesis studies the properties of the lensless optics and proposes the dedicated deep learning algorithms for the encoded pattern processing.

3 | Improved Image Reconstruction

Computational imaging technology is concerned with designing post-capture computational algorithms to form a desired image. Computational imaging can move partial imaging burden from optics to computation. Traditional lensed cameras, where all imaging burden is on optics, have to rely on bulky and complex lens system to achieve bright, sharp and aberration-free imaging. The smartphone camera is a typical application of computational imaging. With compact lens system, the quality of raw captured photos is relatively low. But it finally produces impressive high-quality photos after some post-capture processing. Here, partial imaging burden has been moved from optics to computation. Lensless imaging, where almost all imaging responsibility is carried by the computation, takes the concept of computational imaging to an extreme. In lensless imaging, computationally transforming the captured encoded pattern to an image that faithfully resembles the scene is known as "image reconstruction", shown as Figure 3.1. The improvement of image reconstruction remains a main subject in lensless imaging. This chapter reviews existing image reconstruction methods and proposes the Lensless Imaging Transformer (LIT) that yields superior imaging quality.

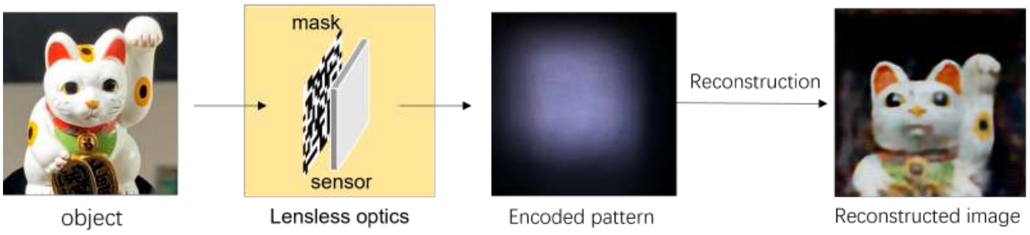


Figure 3.1: Image reconstruction

3.1 Existing Works

There are PSF-based methods and deep neural networks (DNNs) for image reconstruction.

The PSF-based method, also known as the model-based method, is the conventional and the most widely used image reconstruction method. It leverages knowledge of the PSF and models the optical system (Eq. 2.1) as linear by giving the assumption of shift-invariant PSF. The iterative optimization [1, 2, 4, 3], one of the PSF-based methods, solves the target object $\mathbf{o}$ from Eq. 2.1 by iteratively minimizing a loss function [23, 24, 25]:

$$\hat{\mathbf{o}} = \underset{\mathbf{o}}{\text{argmin}} \parallel \mathbf{i_o} - \Phi\mathbf{o} \parallel_{\ell_2} + \tau\Psi(\mathbf{o}), \tag{3.1}$$

where Ψ denotes a regularizer based on an image sparsity prior, and τ is a tuning parameter. Equation 3.1 can be unrolled to a stack of neural network layers to accelerate the reconstruction process [8, 9]. Moiré decoding [6, 7, 26] is another PSF-based method. It decodes the object $\mathbf{o}$ from Eq. 2.1 by using specially designed Fresnel zone apertures with different phases as the mask and decoding with the fringe scanning [27].

PSF-based methods rely on the PSF and need the shit-invariant PSF assumption. However, as stated in Section 2.1, the ideal PSF is essentially impossible to acquire in practice, and the shift-invariant assumption does not always hold. Additionally, the sparsity prior used in the iterative optimization method is not matched to all scenes. These facts result in serious errors and artifacts during reconstruction in PSF-based methods. Though perceptual quality enhancement methods can be applied after reconstruction [8, 9], they can only partially correct the errors and artifacts.

Therefore, designing pure data-driven DNNs for image reconstruction in mask-based lensless imaging is attractive. Pure DNN needs no PSF thus avoiding errors and artifacts caused by it. Besides, DNNs can learn stronger scene priors than the single sparsity prior. Existing pure DNNs for image reconstruction are all based on fully convolutional networks (FCNs) [28, 29, 30]. However, FCNs have limitations as the multiplexing property (Section 3.2) is not considered.

3.2 Multiplexing Property

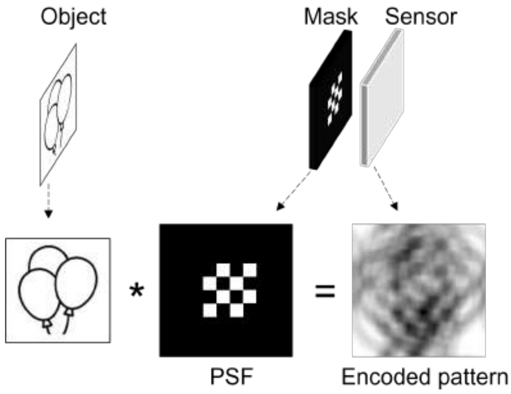


Figure 3.2: Example of multiplexing property

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. It also means that each sensor pixel measures multiplexed light from widely spread points in the scene. This is known as the multiplexing property in the mask-based lensless optics. The multiplexing property transforms local information in the scene to overlapping global information in the image sensor. A simulation example is illustrated in Figure 3.2 where the object is a two-dimensional (2D) object, the mask is a 4×4 binary amplitude mask and the encoded pattern is simplified to be the result of convolution between the object and the PSF.

Consequently, global feature extraction is essential for encoded pattern understanding. The FCN is built on stacking, weight-sharing and small-kernel convolutions (such as 3×3 or 7×7) in a deep network [31, 32, 33, 34, 35, 36]. Such design takes the advantage of locality prior for natural images. It is efficient in reasoning local features such as edges, ridges and corners which focus on a small group of pixels, but its access to global features is limited because it is only done in deeper layers. Meanwhile, long range dependencies are essential in the encoded pattern considering the multiplexing property. Based on this consideration, this thesis argues that FCNs may not be the best choice for the lensless camera despite their domination in most imaging processing tasks.

3.3 Lensless Imaging Transformer (LIT)

Based on the discussion in Section 3.2, a Transformer-based neural network for pure data-driven image reconstruction, dubbed as the "lensless imaging Transformer", is proposed. The Transformer is free of the inductive bias of locality and considers global context straightly by encoding with self-attention mechanism, thus enhancing global feature reasoning [37, 38, 39]. The diagrammatic overview of proposed LIT is illustrated in Figure 3.3. Standardization, which rescales the data to make the mean and the standard deviation be 0 and 1, respectively, is performed for the input encoded pattern. Considering that the reconstruction should be identical for each color channel, the network receives and outputs one channel. For a colorful image, R, G, B channels are reconstructed separately and then concatenated. Following SegFormer [40], 4 encoder blocks are laid to learn hierarchical feature representations from the input. Hierarchical feature representations are then upsampled to the target size, concatenated and decoded.

Each encoder block consists of the overlapped patchify and the Transformer. The

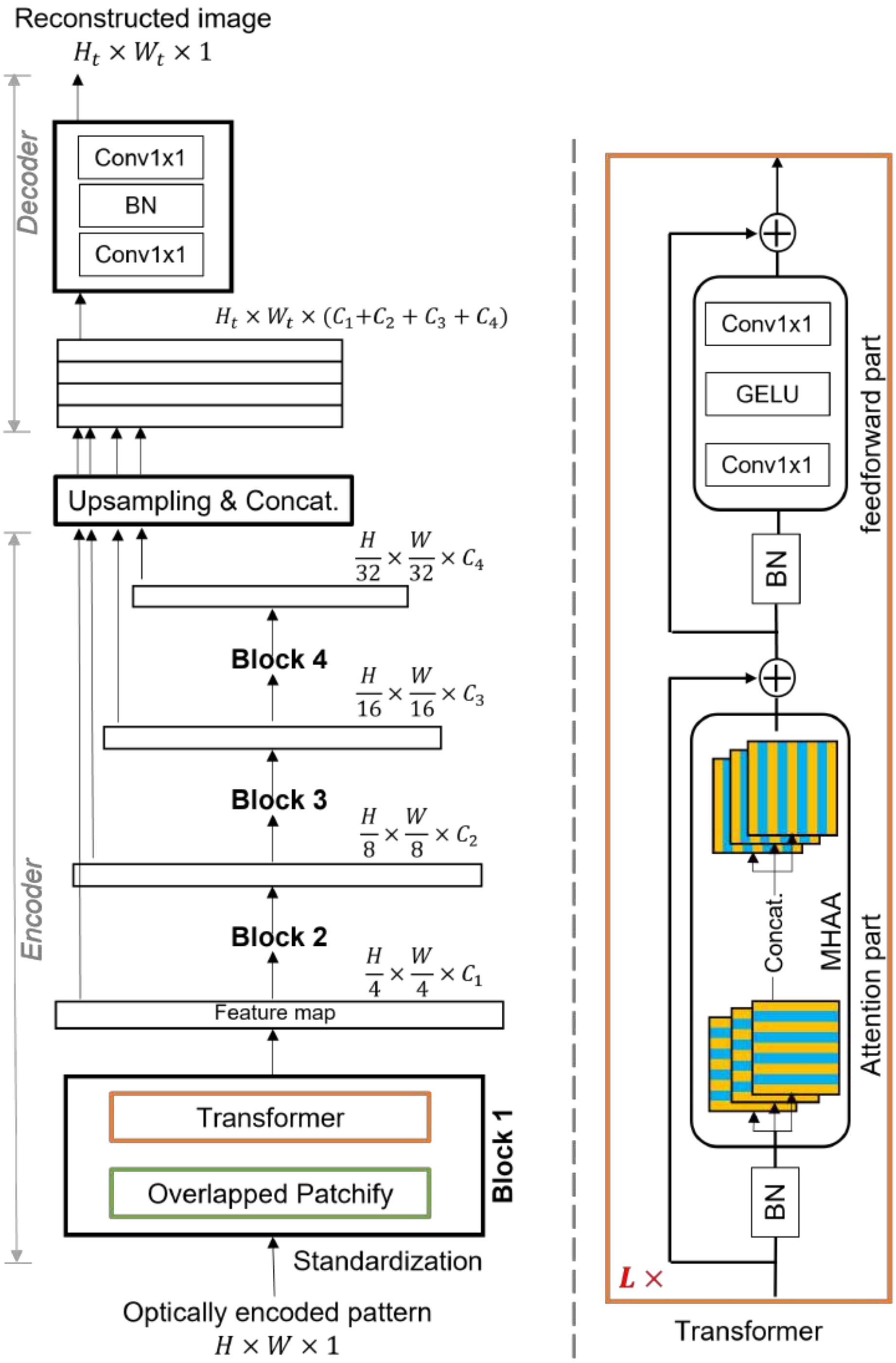

Reconstructed image
$H_t \times W_t \times 1$
Decoder
Conv1x1
BN
Conv1x1
$H_t \times W_t \times (C_1+C_2+C_3+C_4)$
Upsampling & Concat.
$\frac{H}{32} \times \frac{W}{32} \times C_4$
Block 4
$\frac{H}{16} \times \frac{W}{16} \times C_3$
Block 3
$\frac{H}{8} \times \frac{W}{8} \times C_2$
Block 2
$\frac{H}{4} \times \frac{W}{4} \times C_1$
Feature map
Encoder
Transformer
Overlapped Patchify
Block 1
Standardization
Optically encoded pattern
$H \times W \times 1$
Conv1x1
GELU
Conv1x1
feedforward part
BN
Concat.
MHAA
Attention part
BN
$L \times$
Transformer

Figure 3.3: Overview of LIT

Table 3.1: Variants in the encoder of LIT

	Patchify				Transformer	
Block No.	K	S	P	C	L	N
1	7	4	3	64	2	1
2	3	2	1	128	2	2
3	3	2	1	256	2	4
4	3	2	1	512	2	8

overlapped patchify reduces the size and increases the depth of the input while the Transformer maintains the input shape. The overlapped patchify applies *K*-kernel size, *S*-stride, *P*-padding and *C*-number convolutions to extract overlapping embedded patches. *K*, *S*, *P* and *C* for each block are listed in Table 3.1. The overlapped patchify can preserve local continuity around patches which is beneficial for learning pixel-level transformation. The computation bottleneck is the Transformer. To be applicable for large input size, computation reduction in the Transformer is needed. Thus replacing the traditional self-attention by the axial-attentions [41] is proposed. The proposed Transformer has *L* layers. *L* for each block is listed in Table 3.1. Each layer 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 (MHAA), as shown in the right part of Figure 3.3. The feedforward part includes BN, stride-(1×1), kernel-(1×1), number-d_{en} convolutions, Gaussian error linear unit (GELU) and stride-(1×1), kernel-(1×1), number-D convolutions in order. D is set as same as the input depth and d_{en} is set as $4D$.

In MHAA, N axial-attentions are computed in parallel and finally concatenated. N for each block is listed in Table 3.1. Axial-attentions consequently performs self-attention on the feature map width and height axes, formally:

$$y_o = \sum_{p \in N_{1 \times m(o)}} softmax_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)(v_p + r_{p-o}^v). \tag{3.2}$$

y_o is the output at position o. Queries $\mathbf{q} = \mathbf{W_Q}\mathbf{i}$, keys $\mathbf{k} = \mathbf{W_K}\mathbf{i}$, and values $\mathbf{v} = \mathbf{W_V}\mathbf{i}$ are linear projections of the input $\mathbf{i} \in R^{H_i \times W_i \times C_i}$, where $\mathbf{W_Q} \in R^{\frac{C_i}{2N} \times C_i}$, $\mathbf{W_K} \in R^{\frac{C_i}{2N} \times C_i}$ and $\mathbf{W_V} \in R^{\frac{C_i}{N} \times C_i}$ are learnable matrices. $r^q_{p-o}, r^k_{p-o}, r^v_{p-o} \in R^{m \times m}$ 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. $m = W_i$ for the width-axial attention, and $m = H_i$ for the height-axial attention. Unlike traditional self-attention which processes the data as a 1D sequence, axial-attentions regards the data as 2D graphics. Taking advantage of geometry construction information, axial-attentions reduces the computational complexity by processing pixels in width and height axes separately. Replacing self-attention with axial-attentions reduces the computational complexity from $\mathcal{O}(H_i^2 W_i^2)$ to $\mathcal{O}(H_i W_i m)$.

The decoder is a feedforward network including stride-(1×1), kernel-(1×1), number-d_{de} convolutions, BN and stride-(1×1), kernel-(1×1), number-1 convolution in order. d_{de} is 256.

3.4 Experiments

The proposed LIT, FCN, and the PSF-based method are compared. For FCN, U-net [35] is implemented. U-net architecture is outlined in Table 3.2 where "up()" and "," represent bilinear upsampling and concatenation, respectively. Each encode is the stride-(3×3), kernel-(1×1) convolution followed by BN, ReLu [42] and the stride-(2×2), kernel-(2×2) pooling. Each decode is the stride-(3×3), kernel-(1×1) convolution followed by BN and ReLu. For the PSF-based method, the iterative optimization method (Eq. 3.1), which is considered the most representative, is used. The iterative optimization employs alternating direction method of multipliers (ADMM) [25]; the used regularization is total-variation [43] and the tuning parameter τ is set as 10^{-4}.

Table 3.2: Architecture of U-net used for comparison

encode	cha. in/out	input	decode	cha. in/out	input
enc1	1/64	input	dec5	1024/512	up(conv1),enc5
enc2	64/128	enc1	dec4	512/256	up(dec5),enc4
enc3	128/256	enc2	dec3	256/128	up(dec4),enc3
enc4	256/512	enc3	dec2	128/64	up(dec3),enc2
enc5	512/1024	enc4	dec1	64/64	up(dec2),enc1
conv1	1024/1024	enc5	conv2	64/1	dec1

3.4.1 Optical Setup of Lensless Camera

For the optical experiments, a mask-based lensless camera is assembled, shown as Figure 3.4. The mask is placed 2.5 mm in front of the image sensor. Since the objective is evaluating reconstruction algorithms, a pseudorandom binary amplitude mask without any particular optimization is used. The mask is fabricated by chromium deposition in a synthetic-silica plate, with aperture size of 40×40 µm. The sensor is a 6.41 megapixels CMOS (Sony IMX178) with 2.4×2.4 µm pixel pitch, and 12 bit color depth. The PSF is captured by illuminating the mask with a point LED with the diameter of 1 mm placed 15 cm away. The PSF is only used in the PSF-based method.

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

$$Resolution = \frac{d_1}{d_2} \times s, \tag{3.3}$$

where d_1, d_2 and s correspond to the 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 approximately 0.144

mm when the target is 15 cm away. The degree of the diffraction effect can be evaluated by the Fresnel number N_F [22]

$$N_F = \frac{a^2}{L\lambda}, \tag{3.4}$$

where a is the mask aperture size, L is the mask-sensor distance and λ is the incident wavelength. Here, $a = 40$ µm, $L = 2.5$ mm, and λ is 380∼700 nm for visual light. Considering the incident wavelength to be 500 nm, the calculated N_F is 1.28, larger than 1, which means the diffraction effect is well constrained in relation to the geometric size.

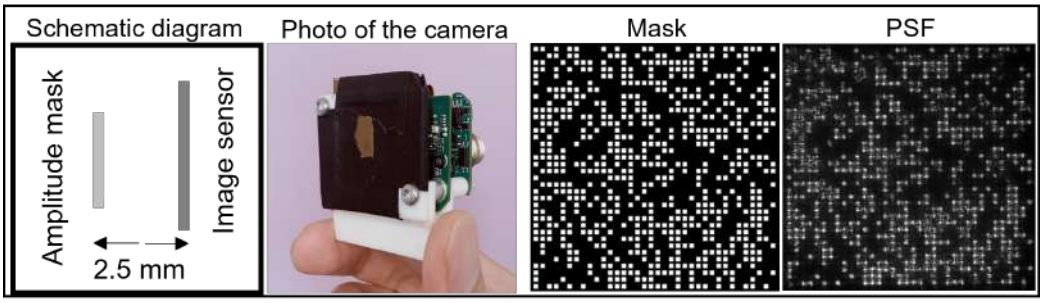


Figure 3.4: Assembled Lenselss Camera

3.4.2 Experiment Details

The optically encoded pattern dataset is collected by shooting while displaying images on an LCD screen placed around 15 cm ahead the lensless camera. The original image dataset consists of MirFlickr dataset [44] (25,000 RGB images) and a part of ILSVRC-2012 ImageNet dataset [45] (35,000 RGB images are selected). One thousand images are selected for validation and others for training. All displayed images are resized to 500×500 pixels, and captured encoded patterns occupy within 1,600×1,600 pixels on the sensor.

In the LIT, the input and output size are set as 1,600×1,600 and 500×500 pixels, respectively. Both input and output size of the U-net are set as 1,600×1,600 pixels. In training, the input is the captured encoded pattern and ground truth is the resized original image. Loss is the mean squared error (MSE) between the output and the corresponding ground truth image which is resized to 500×500 pixels for the LIT and 1,600×1,600 pixels for the U-net. Both networks are trained on a machine equipped with an Intel Xeon E5-2698 v4 CPU (2.2 GHz), two NVIDIA TESLA V100 GPUs (2 × 32 GB), Python 3.6.5, and Pytorch 1.7.1. The optimizer employs Adam [46] with β_1=0.9 and β_2=0.999. The weight decay is 0.1. With a batch size of 4, learning rate of 6×10^{-5}, learning rate warmup of 15,000 steps and cosine learning rate decay, the LIT and U-net converge within 250,000 and 150,000 steps, respectively.

3.4.3 Results

Figure 3.5 and 3.6 show the experimental results with on-screen images and in-wild objects, respectively. Both the screen and objects are placed around 15 cm ahead of the camera. The iterative optimization method reconstructs a 1,600×1,600 pixels encoded pattern into a reconstructed image with the same size. The ADMM runs 50 iterations ensuring convergence. Running on the Intel Xeon E5-2698 v4 CPU, the proposed method and U-net take 0.9s and the iterative optimization method takes more than one hour for one image reconstruction. Peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM) are calculated for the on-screen experiment. PSNR/ SSIM of the downy cat image is 29.18dB/ 0.54 for the proposed method and 28.76dB/ 0.34 for the U-net. PSNR/ SSIM of the house image is 28.40dB/ 0.56 for the proposed method and 27.61dB/ 0.21 for the U-net.

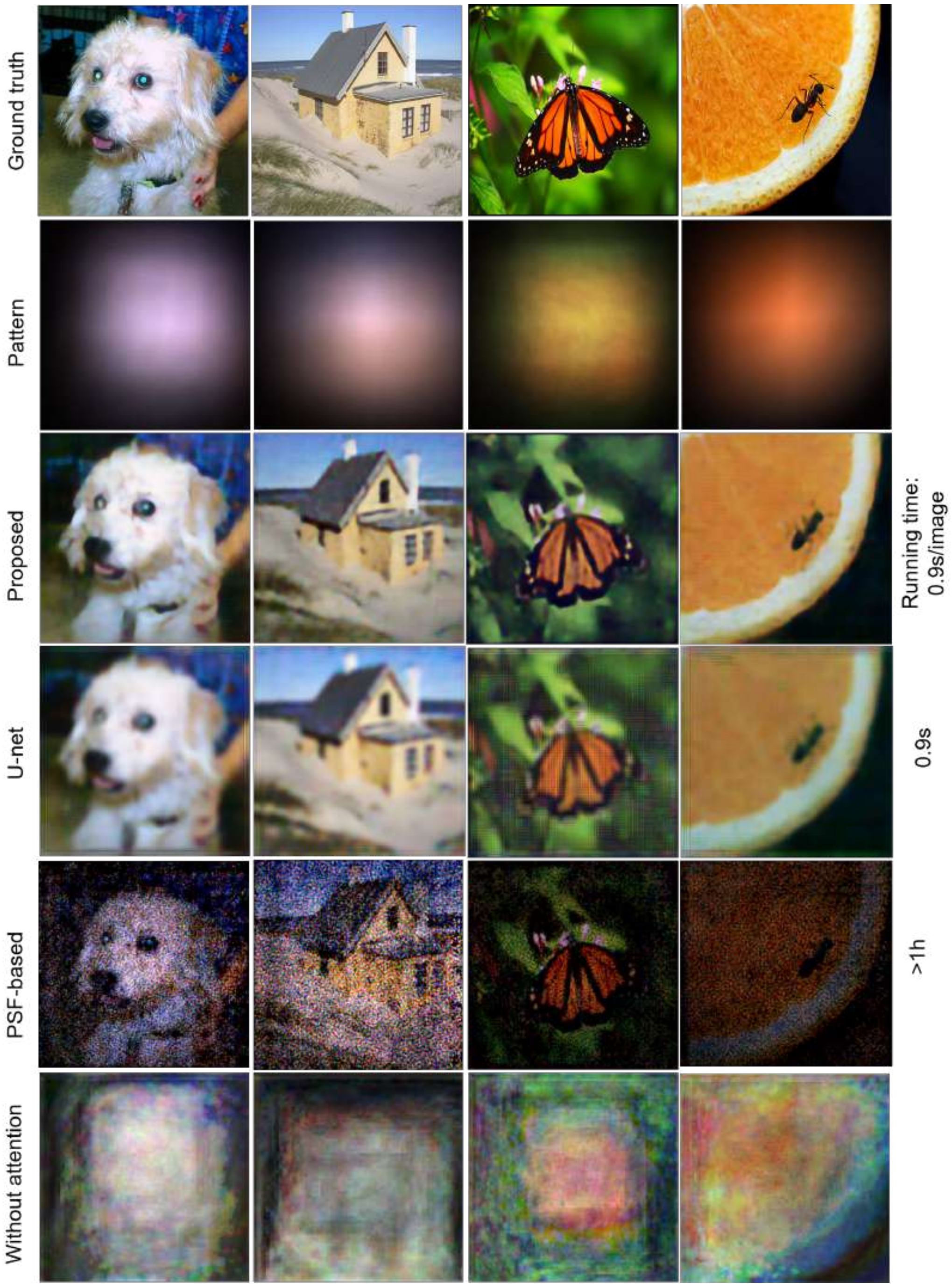


Figure 3.5: Experimental results with on-screen samples

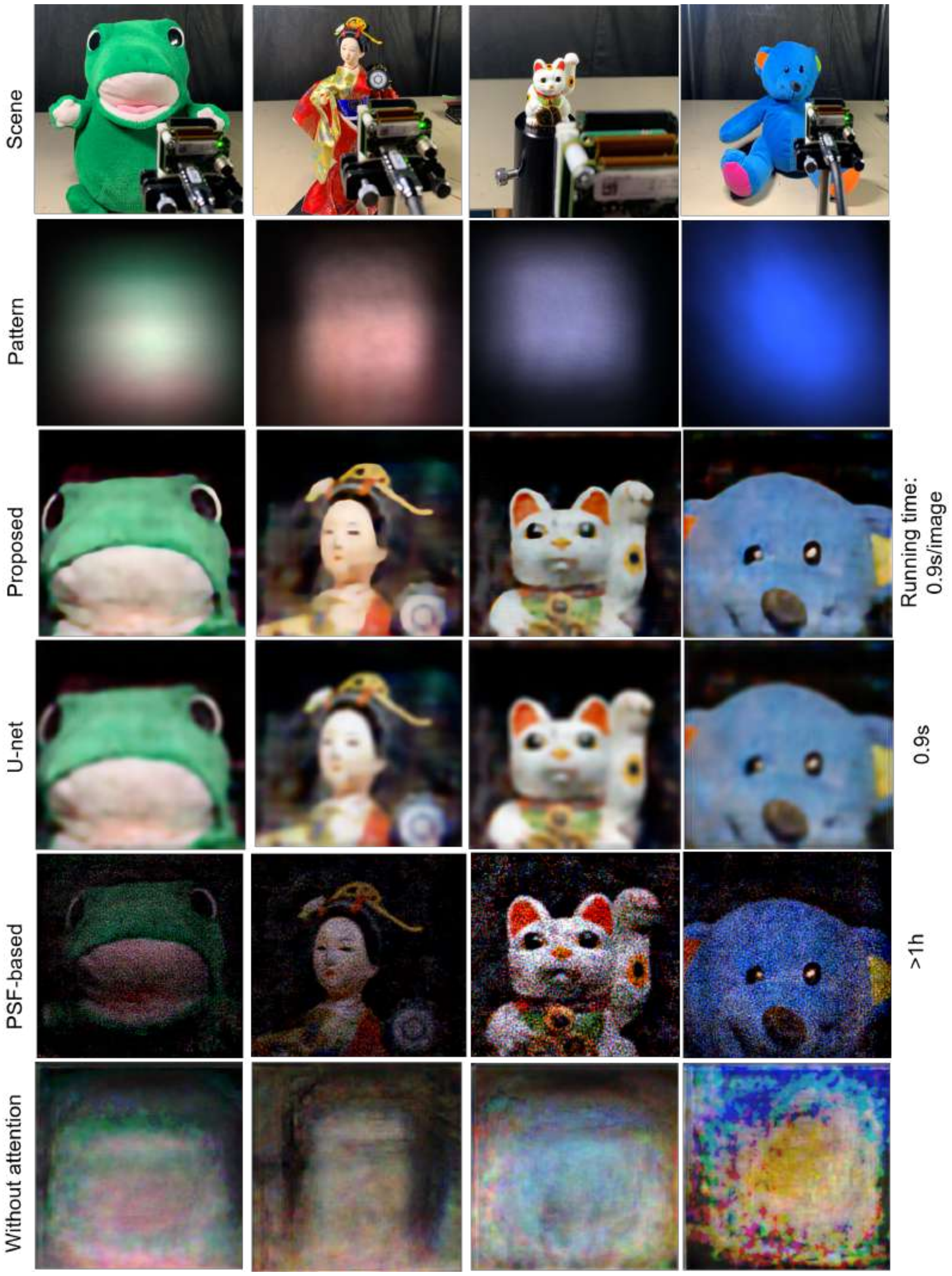


Figure 3.6: Experimental results with in-wild samples

The results show that the U-net reconstructs images with lacking details compared with the reconstructed images by the proposed method. The PSF-based method can reconstruct images with a similar resolution as the proposed method does. But the PSF-based method is unable to reconstruct correct color in some areas and it introduces evident streaky artifacts and noises. These factors diminish the visual quality of the reconstructed images. Overall, the proposed method produces the most visually appealing images.

An ablation study on the Transformer component in the proposed network is also performed. By deleting the attention part in the Transformer component, the model with the same training conditions fails to reconstruct meaningful images, as shown in the last row of Figure 3.5 and 3.6. It verifies the main contribution of the Transformer in the proposed network.

3.5 Discussion

This chapter studies the reconstruction algorithm which is the most fundamental part in the lensless imaging. The proposed LIT is proven superiority compared with the FCN and the PSF-based method.

There is trade-off between performance and model size in designing LIT. The model size is related to the computation complexity, difficulty of training and amount of demanded training data. For reduced computation and ease of training, LIT makes two sacrifices. Firstly, LIT applies convolutional layers (overlapped patchify) to learn features with reduced size in ahead of Transformer. This design greatly simplifies the model, but may not be optimal way to learn global features. Secondly, LIT receives color channels one by one and processes them as the same, ignoring the difference of diffraction effect and information relations between channels. Designing the model to better optimize this trade-off is a valuable future work.

There are two points this thesis recommends considering when applying the proposed reconstruction algorithm. First, pure data-driven methods, including the proposed one, are restricted by the training data distribution. Learned image priors greatly rely on the fed training data. The domain difference between the training and application data should not be too large. Second, optimization of mask design and adding perceptual quality enhancement or denoise algorithm afterwards could further improve the image quality.

Though the proposed algorithm is designed for the mask-based lensless imaging, it could potentially extended to more applications such as image reconstruction in mask-free lensless imaging [47], through-scattering imaging [48] and under-screen imaging.

4 | Reconstruction-free Recognition

With the recent proliferation of artificial intelligence, in many situations, the ultimate goal of visual information acquisition has been transferred from human visual appreciation to computer vision for machine intelligence. Based on this observation, computer vision could be an integral application of the lensless camera. To apply lensless camera in computer vision, traditional way requires image reconstruction prior to performing computer vision algorithms [49]. This thesis argues that image reconstruction is not intrinsically needed in terms of computer vision tasks. Though the encoded pattern does not seem human-interpretable, it contains sufficient visual information about the target to allow the machine to understand. Image reconstruction aims to recreate a "normal" image that human can understand. Indeed, the machine does not necessarily understand the world as human does, thus image reconstruction is not a must. Image reconstruction not only uses additional computational resources but also introduces errors 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 [50] or de-identified attribute recognition like gender and age estimation. Therefore, this thesis proposes to directly perform object recognition on the encoded pattern, bypassing tedious image reconstruction. Figure 4.1 illustrates the comparison among the proposed reconstruction-free lensless camera, the lensed camera and the reconstruction-including lensless camera considering an object recognition task. The lensless camera is advantageous in optical hardware than the lensed camera. Avoiding additional errors and computation caused by reconstruction, the reconstruction-free lensless recognition is more efficient in both prediction accuracy and computation than the reconstruction-including lensless recognition.

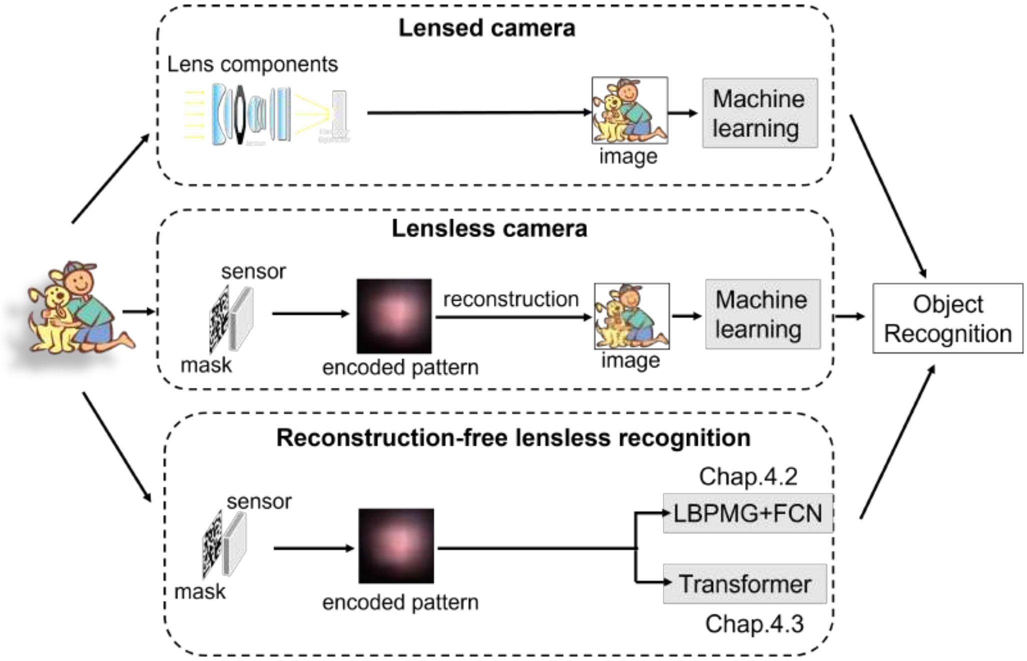


Figure 4.1: Reconstruction-free lensless recognition and other ways

As stated in Section 3.2, there is multiplexing property in the lensless optics and FCNs are insufficient in handling the encoded pattern. Achieving reconstruction-free lensless recognition by directly applying the FCN should be challenging. Here, two solutions are proposed. One involves designing a dedicated computation-efficient data preprocessing method, "Local Binary Pattern Map Generation (LBPMG)", which is introduced in Section 4.2. Inspired by Chapter 3, another solution abandons FCNs and focuses on developing Transformer-based neural network for reconstruction-free lensless recognition. The proposed Transformer-based neural network, "Lensless Recognition Transformer (LRT)", is introduced in Section 4.3.

4.1 Related Works

Reconstruction-free recognition is attracting significant interest and has been applied across several fields. Speckle patterns produced by coherent laser exposure can provide information for inference tasks with no necessity of image reconstruction [51, 52, 53, 54, 55, 56]. Speckle pattern recognition has been used in biomedicine [52, 53], agricultural crops investigation [54] and non-line-of-sight recognition [55, 56]. Action recognition without image reconstruction in compressive video sensing gains advantages of privacy-preserving [57] and computation reduction [58]. Reconstruction-free recognition has also been introduced to the single-pixel camera to reduce measurement rates or save computation and sensor pixels [59, 60, 61, 62, 63].

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

4.2 LBPMG with FCN

The multiplexing property transforms local information in the scene to global information in the encoded pattern. Similarly, the multiplexing property amplifies local disturbance to serious global noise. Alleviating the disturbance amplification issue could benefit encoded pattern analysis. This section first discusses the disturbance amplification issue (Section 4.2.1) and then proposes LBPMG data preprocessing method to alleviate the disturbance amplification issue (Section 4.2.2).

4.2.1 Disturbance Amplification Issue

A small local disturbance $\mathbf{d}$ becomes the global noise $\mathbf{i_d}$ through the mask, formulated as

$$\mathbf{i_d} = \Phi\mathbf{d}. \tag{4.1}$$

Figure 4.2 illustrates the disturbance amplification issue, where the mask is a 4×4 binary amplitude mask and the encoded pattern is simplified to be the result of convolution between the scene and PSF. The transformation system is heavily ill-conditioned because the kernel size of the PSF is large compared with the object. The center column of Figure 4.2 shows the local disturbance $\mathbf{d}$ and amplified noise $\mathbf{i_d}$. From the top row of Figure 4.2, we can observe that the local disturbance is small enough that it causes little difficulty for recognizing the object. Conversely, the local disturbance through the mask becomes global noise $\mathbf{i_d}$ overlaying on sensor measurement of the object $\mathbf{i_o}$, as demonstrated at the bottom row of Figure 4.2. It significantly hinders object recognition.

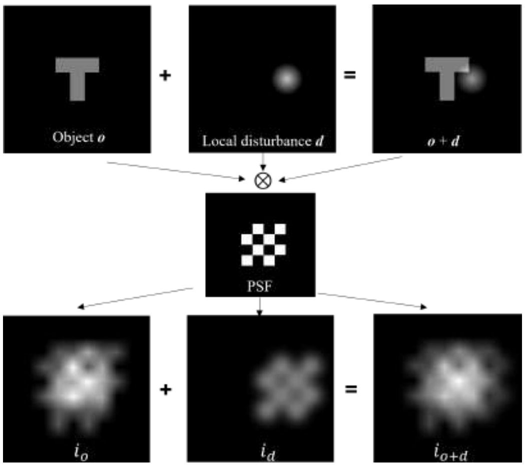


Figure 4.2: Disturbance amplification issue

4.2.2 Local Binary Pattern Map Generation

Note that the encoded pattern shares some similarities with the textured image. For example, to indicate an object in the encoded pattern or textured image, the way the pixels are arranged tends to be a more important characteristic than the shape. The local binary patterns histogram (LBPH) algorithm is the mostly used texture descriptor [64, 65]. The LBPH algorithm represents each image pixel with a binary code and the image is described by the distribution of binary codes. The LBPH algorithm is widely used in texture analysis because it has a low computation cost and is invariant to grayscale illumination changes. Inspired by this, this thesis proposes to generate a 2D map, based on the LBPH algorithm, for the encoded pattern in order to improve the robustness to disturbances. Following the LBPH algorithm, firstly the binary code of each pixel in the encoded pattern is built considering the differences between the pixel and its equally spaced neighbors. Taking the value of the center pixel as the threshold, neighbor pixels are assigned with new binary values by setting 1 for values equal to or higher than the threshold and 0 for values lower than the threshold. These binary values are concatenated into a binary code. An illustration of this process is depicted in Figure 4.3, where the binary code is calculated by thresholding a 3×3 neighborhood. To describe texture feature, the LBPH algorithm then generates a histogram by calculating the distribution of binary codes. Here, the LBPH algorithm is employed as the data preprocessing rather than feature extraction. Hence, I do not generate the histogram but an LBP map by converting the binary code of each pixel to a decimal value. Note that calculating the LBP map is computationally simple enabling real-time image analysis.

Since the binary code of a pixel is generated by comparing it with its surrounding pixels, an LBP map is a more robust format than the encoded pattern in terms of

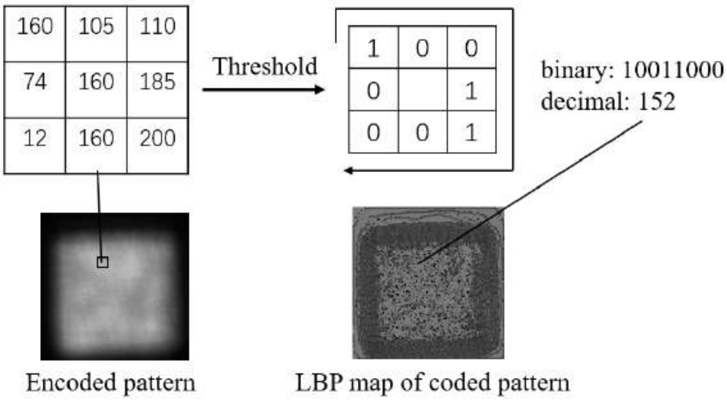


Figure 4.3: The process of LBPMG

handing disturbances. An example of LBP map and its efficiency of disturbance suppression is shown as Figure 4.4. The LBP map of encoded pattern is then used as input for the FCN.

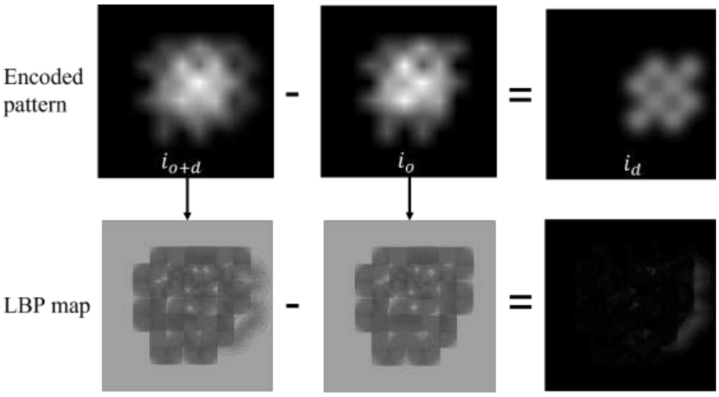
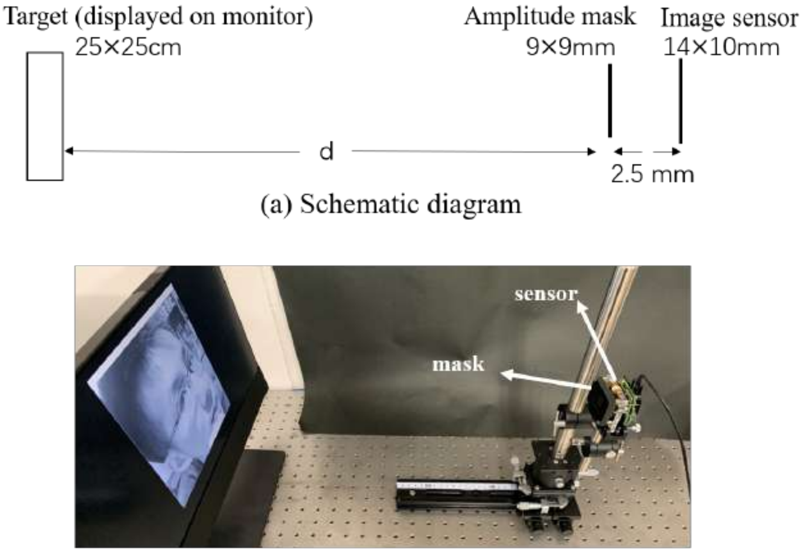


Figure 4.4: LBP map and its efficiency in disturbance suppression

4.2.3 Experiments

To analyze the performance of LBPMG, a series of optical experiments are conducted. The setup of optical experiment is shown as Figure 4.5. The monitor, mask and sensor are kept parallel, and their center points are lined up. The target is displayed on the center of a monitor placed in front of the lensless camera.

The displayed image has a resolution of 200×200 pixels and physical size of 25×25 cm. The binary amplitude mask, fabricated with chromium deposition in a synthetic-silica plate, presents a binary pseudorandom array. Shown as Figure 4.6, the mask has an optical size of 9×9 mm and aperture size of 30×30 µm. The ratio of transmission area is 25%, and each transmission spot occupies 6×6 µm. The sensor is a monochrome 12.3 megapixels CMOS (Sony IMX304) with a 14×10 mm optical size, and 3.45×3.45 µm pixel size. The image senor and mask are separated by 2.5 mm, capturing approximately 75° field-of-view. With Eq. 3.3, the object plane resolution is around 0.0014 **d**, where **d** is the distance between the target and mask. The object plane resolution is 0.49 mm when **d** is 35 cm.



(a) Schematic diagram

(b) Photography of the setup

Figure 4.5: Experiment setup

Definition and explanation of datasets are illustrated in Figure 4.7. The dataset, displayed on the monitor, is named as normal dataset (NDS). NDS contains the normal training dataset (NTDS) and normal validation dataset (NVDS). The dataset,

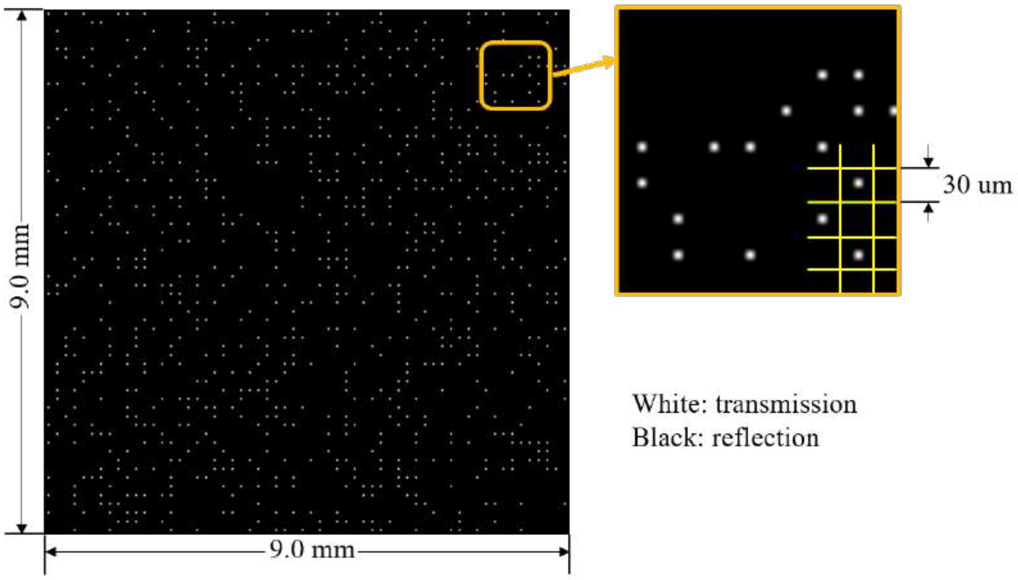


Figure 4.6: Binary amplitude mask

consisting of encoded patterns recorded on the sensor, is named as encoded dataset (EDS). EDS contains the encoded training dataset (ETDS) and encoded validation dataset (EVDS) corresponding to the NTDS and NVDS, respectively. To compare reconstruction-free lensless camera with the reconstruction-including lensless camera, I build the reconstructed dataset (RDS), containing the reconstructed training dataset (RTDS) and reconstructed validation dataset (RVDS), by performing image reconstruction on the EDS. The used image reconstruction algorithm is the iterative optimization (Eq. 3.1) which employs ADMM and total-variation and the tuning parameter τ is set as 10^{-4}. Figure 4.8 illustrates the reconstructed images with different iterative counts. Reconstructed images with 1 iterative count and 10 iterative counts will be used in RDS.

Experiments on handwritten digit recognition and gender estimation are performed. For handwritten digit recognition, NDS uses the MNIST dataset [66] which contains 60,000 training images for NTDS and 10,000 validation images for NVDS. For gender estimation, around 60,000 cropped female faces and 60,000 cropped

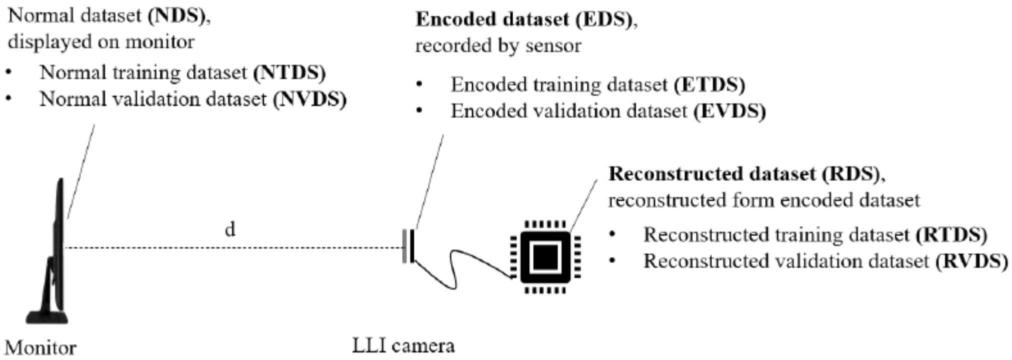

Normal dataset (NDS),
displayed on monitor
• Normal training dataset (NTDS)
• Normal validation dataset (NVDS)
Encoded dataset (EDS),
recorded by sensor
• Encoded training dataset (ETDS)
• Encoded validation dataset (EVDS)
Reconstructed dataset (RDS),
reconstructed form encoded dataset
• Reconstructed training dataset (RTDS)
• Reconstructed validation dataset (RVDS)
d
Monitor
LLI camera

Figure 4.7: Definition and explanation of datasets

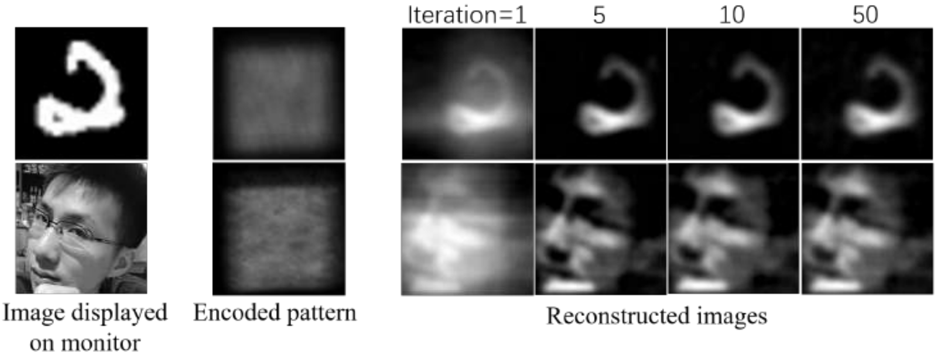


Figure 4.8: Reconstructed images by ADMM

male faces are collected from the IMDB-WIKI dataset [67] for the NTDS. The LFW dataset [68] which is a benchmark dataset for gender estimation, is used for NVDS. Since face images from both the IMDB-WIKI and LFW datasets are captured with varying position, pose, lighting, expression, background, camera quality, occlusion and age, gender estimation with IMDB-WIKI and LFW datasets is a more difficult and advanced inference task than handwritten digit recognition with the MNIST dataset where all digits are white and written on a black background without noise.

LBPMG uses a 3×3 thresholding neighborhood. The FCN applied for all experiments is ResNet-18 [36], whose input layer is modified to receive 224×224 grayscale images. Thus, all images from NDS, EDS and RDS are resized and transformed to 224×224 grayscale. For all tasks, ResNet-18 uses Adam with β_1=0.9, β_2=0.999, a weight decay of 0.1, and a mini-batch size of 64. The training is conducted on two NVIDIA GeForce GTX TITAN X GPUs, with Keras 2.3.1, Tensorflow 1.2.1 and Python 2.7. Training on NDS, EDS and RDS take approximately the same time. For the handwritten digit dataset, the model converges within 5 epochs and each epoch takes approximately 4 minutes. For the gender dataset, the model converges within 8 epochs and each epoch takes approximately 12 minutes.

Experiments under uneven illumination (Experiment 1, Section 4.2.3.1) and with a moving target (Experiment 2, Section 4.2.3.2) are performed.

4.2.3.1 Experiment 1

In Experiment 1, the monitor is fixed 35 cm away from the lensless camera. There are many disturbances in real environment. I choose uneven illumination, one of the most common disturbances, for evaluation. To simulate uneven illumination, I beforehand create normal dataset with local illumination variation (NDS-IV) by randomly adding local illumination variation to the NDS. NDS-IV contains the normal training dataset with local illumination variation (NTDS-IV) and normal validation dataset with local illumination variation (NVDS-IV). The effect of adding local illumination variation is shown in Figure 4.9. Note that face images from the IMDB-WIKI and LFW datasets are captured in the wild where changes in lighting conditions already exist. It means that, for gender estimation, both NDS/ RDS/ EDS and NDS-IV/ RDS-IV/ EDS-IV contain illumination variations, but NDS-IV/ RDS-IV/ EDS-IV contain more.

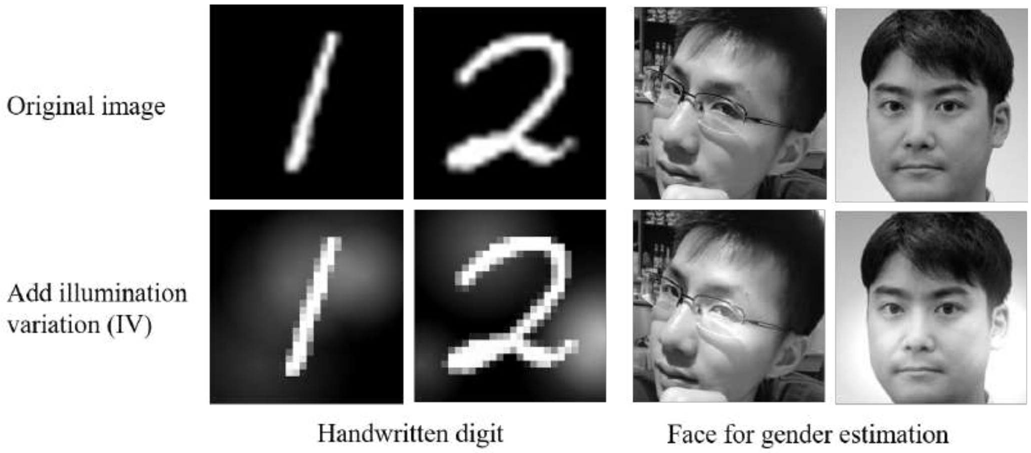


Figure 4.9: Effect of randomly adding illumination variation in local

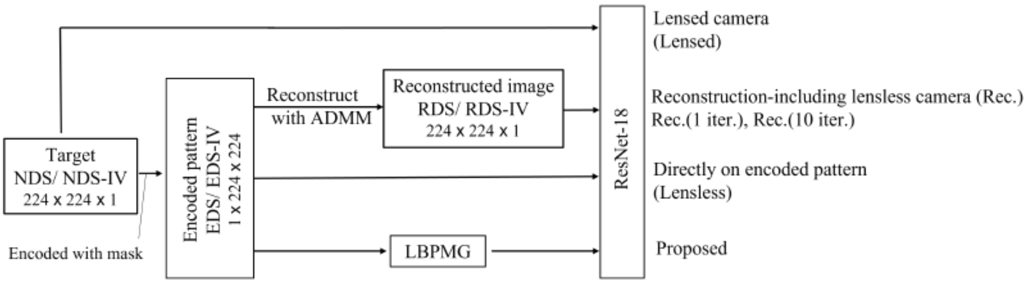


Figure 4.10: Experiment schemes of Experiment 1

Experimental schemes of Experiment 1 are illustrated in Figure 4.10. Note that in all schemes, normalization is conducted before images are sent to ResNet-18. Predictive accuracy result is listed in Table 4.1. In Table 4.1, the lensed camera is abbreviated as "lensed", the reconstruction-including lensless camera as "Rec.(1 iter.)" for image reconstruction with 1 iterative count and "Rec.(10 iter.)" for 10 iterative counts, the directly recognition on the encoded pattern as "Lensless", and the proposed "LBPMG + FCN" is abbreviated as "LBPMG". "Digit" and "gender" represent handwritten digit recognition and gender estimation, respectively. "IV" indicates illumination variation. Computation speed comparison is listed in Table 4.2, where all calculations are timed with a 224×224×1 image in Windows 10, Keras 2.3.1, Tensorflow 1.2.1 and Python 2.7, on a machine with a CPU i7-7700.

Table 4.1: Predictive accuracy results of Experiment 1

Test #	Training/ Val.	Lensed	Rec.(1 iter.)	Rec.(10 iter.)	Lensless	LBPMG
Digit						
D1	no IV/ no IV	99.12%	97.11%	95.55%	98.83%	98.87%
D2	no IV/ with IV	99.37%	96.55%	93.42%	11.35%	60.81%
D3	with IV/ with IV	98.76%	96.59%	94.93%	89.53%	97.74%
Gender						
G1	no IV/ no IV	89.54%	65.57%	69.50%	50.52%	86.26%
G2	no IV/ with IV	84.08%	64.98%	69.37%	50.40%	77.29%
G3	with IV/ with IV	88.38%	65.48%	69.43%	61.45%	77.83%

Table 4.2: Computation speed comparison

	LBPMG	ResNet-18	Rec. with 1 iter.	Rec. with 10 iter.
Time consumed	0.16 ms	0.96 s	1.82 s	9.60 s

Comparing direct recognition on the encoded pattern and the proposed "LBPMG + FCN" in Table 4.1, there are two observations. First, direct recognition on the encoded pattern is possible for simple inference tasks like handwritten digit recognition in controlled lighting conditions (test D1, 98.83%). Whereas it is not feasible (test D2, 11.35%) or has poor performance (test D3, 89.53%) when uneven illumination is added to the target. The reason is the disturbance amplification issue, as explained in Section 4.2.1. For advanced task like gender estimation, direct recognition on the encoded pattern becomes infeasible (test G1, 50.52%; G2, 50.40%; G3, 61.45%). The second observation is that LBPMG is an efficient data preprocessing method to suppress noise and benefits the predictive accuracy. With LBPMG, it achieves a high predictive accuracy for simple tasks like handwritten digit recognition no matter without uneven illumination (test D1, 98.87%) or with uneven illumination (test D3, 97.74%). For advanced tasks like gender estimation, it achieves a decent accuracy even when uneven illumination exists.

Now we compare predictive accuracy among the proposed "LBPMG + FCN", the lensed camera and the reconstruction-including lensless camera in Table 4.1. the proposed one surpasses reconstruction-including lensless camera, and achieves an accuracy that is close to the lensed camera's in both conditions without and with uneven illumination. As for computational efficiency, as shown in Table 4.2, LBPMG consumes 0.16 millisecond, which can be ignored when compared to the inference with ResNet-18 or iterative image reconstruction with ADMM.

4.2.3.2 Experiment 2

To further evaluate LBPMG in a real environment, I change 3D positions of the target during the experiment. The experimental setup is shown in Figure 4.11. The origin of the axis is set in the center of the monitor, 35 cm away from the camera. Translation in z direction is accomplished by moving the monitor, and translation in the x or y directions is accomplished by moving the area of the display window on the monitor. By changing the target's 3D position (x,y,z), ETDS-(x,y,z) and EVDS-(x,y,z) are collected. ETDS-(x,y,z) or EVDS-(x,y,z) means that the ETDS or the EVDS is collected with the target at the position (x,y,z). In total, 9 ETDS are collected, as listed in Table 4.3. Unit for measurement is centimeter. Model trained with only ETDS-(0,8,8) and model trained with all 9 ETDS are used to test 24 EVDS. Results are shown in Figure 4.12 for handwritten digit recognition and Figure 4.13 for gender estimation. Data is list in Table 4.4.

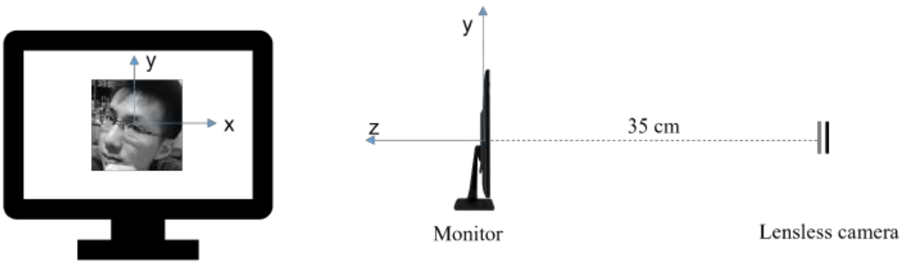


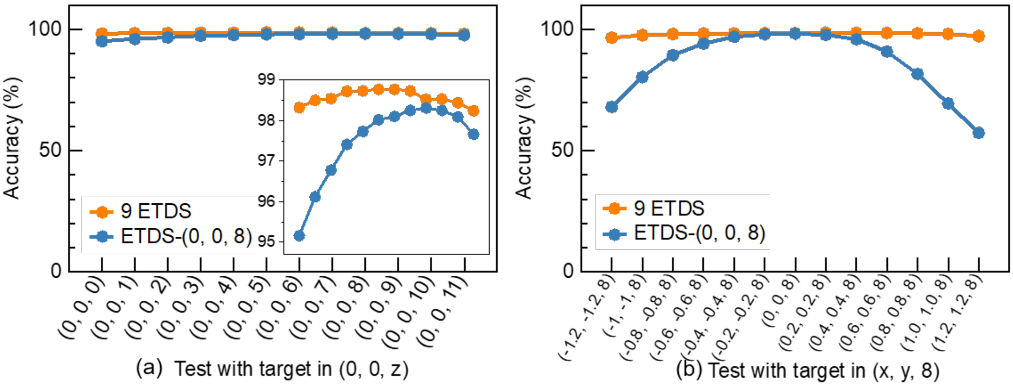
Figure 4.11: Experiment 2 setup

Table 4.3: Training datasets collected in Experiment 2

ETDS-(0, 0, 0)	ETDS-(0, 0, 2)	ETDS-(0, 0, 4)	ETDS-(0, 0, 6)
ETDS-(0, 0, 8)	ETDS-(-0.8, -0.8, 8)	ETDS-(-0.4, -0.4, 8)	ETDS-(0.4, 0.4, 8)
ETDS-(0.8, 0.8, 8)			

For handwritten digit recognition when the target moves in the z direction, shown as Figure 4.12 (a), it keeps a high accuracy, even only ETDS-(0,8,8) is used for training. For the target moving in the x and y directions, shown as Figure 4.12 (b), accuracy decreases dramatically as the target moves away from the training position (0,8,8) if only ETDS-(0,8,8) is used for training. In contrast, accuracy can be maintained if 9 ETDS are used for training. Figure 4.12 reveals that, for handwritten digit recognition, it is robust to the target's z direction translation and sensitive to translations in the x and y directions if the model is trained with the target in only one position.

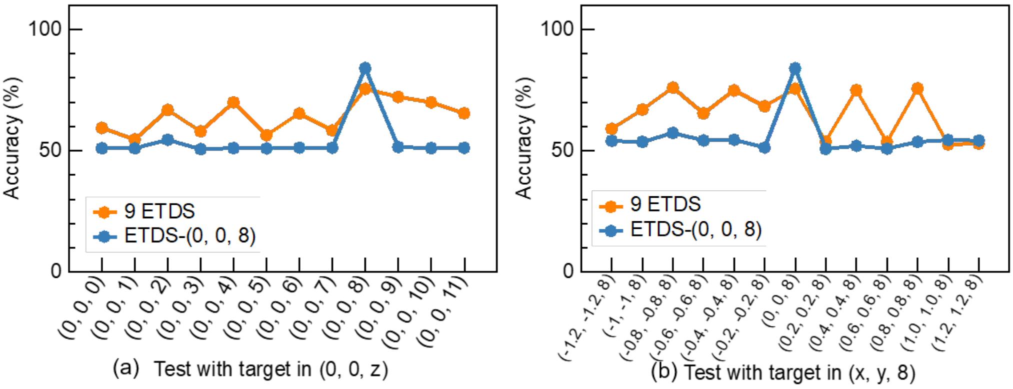
For the more advanced task of gender estimation, shown as Figure 4.13, it has weak robustness to target movement. Conversely, we observe an accuracy improvement when the target position for the test is included in the target positions



(a) Test with target in (0, 0, z)

(b) Test with target in (x, y, 8)

Figure 4.12: Result for handwritten digit recognition with moving target



(a) Test with target in (0, 0, z)

(b) Test with target in (x, y, 8)

Figure 4.13: Result for gender estimation with moving target

Table 4.4: Experiment 2 result

EVDS position	Model trained with ETDS-(0, 8, 8)	Model trained with all 9 ETDS
	digit recognition / gender estimation	digit recognition / gender estimation
(0, 0, 0)	95.16% / 51.00%	98.32% / 59.47%
(0, 0, 1)	96.12% / 51.04%	98.50% / 54.70%
(0, 0, 2)	96.78% / 54.56%	98.54% / 66.82%
(0, 0, 3)	97.00% / 50.67%	98.72% / 58.01%
(0, 0, 4)	97.73% / 51.14%	98.73% / 69.97%
(0, 0, 5)	98.02% / 50.99%	98.77% / 56.47%
(0, 0, 6)	98.10% / 51.33%	98.77% / 65.40%
(0 ,0, 7)	98.25% / 51.31%	98.73% / 58.42%
(0, 0, 8)	98.31% / 84.11%	98.52% / 75.59%
(0, 0, 9)	98.25% / 51.57%	98.53% / 72.22%
(0, 0, 10)	98.09% / 51.13%	98.44% / 70.01%
(0, 0, 11)	97.66% / 51.23%	98.24% / 65.51%
(-1.2, -1.2, 8)	68.09% / 54.13%	96.63% / 59.07%
(-1, -1, 8)	80.43% / 53.67%	97.69% / 67.00%
(-0.8, -0.8, 8)	89.45% / 57.40%	98.14% / 76.07%
(-0.6, -0.6, 8)	94.19% / 54.28%	98.34% / 65.44%
(-0.4, -0.4, 8)	96.99% / 54.56%	98.38% / 74.91%
(-0.2, -0.2, 8)	98.03% / 51.36%	98.53% / 68.39%
(0, 0, 8)	98.31% / 84.11%	98.52% / 75.59%
(0.2, 0.2, 8)	97.78% / 50.83%	98.72% / 53.88%
(0.4, 0.4, 8)	96.02% / 52.04%	98.66% / 75.05%
(0.6, 0.6, 8)	90.93% / 50.91%	98.60% / 53.65%
(0.8, 0.8, 8)	81.69% / 53.71%	98.44% / 75.75%
(1, 1, 8)	69.47% / 54.44%	98.20% / 52.67%
(1.2, 1.2 ,8)	57.38% / 54.23%	97.30% / 52.99%

during training. It means that, for gender estimation, it can maintain a high robustness to target movement when training with sufficient target positions.

The Experiment 2 result actually reveals the generalization issue of deep learning. When the test data deviates too much from the training data domain, deep learning methods fails.

4.3 Lensless Recognition Transformer (LRT)

The proposed "LBPMG +FCN" is a preliminary solution. It is theoretically designed for a cropped and aligned target with a flat background, such as microscopic cells, and does not consider more complicated scenes. Here, inspired by Chapter 3, the end-to-end neural network, LRT is proposed. The LRT is considered a more advanced solution enabling object recognition in complicated backgrounds without cropping or aligning.

Architecture of the LRT is introduced in Section 4.3.1. Notice that, without inductive priors, the Transformer suffers from non-convex losses [21]. It requires a larger training dataset to alleviate this problem. Therefore, I also propose the method to generate simulated encoded patterns for pretraining, which is introduced in Section 4.3.2.

4.3.1 LRT

The LRT primarily comprises the patchify stem and the Transformer. The patchify stem reshapes the input image to non-overlapping patches, which are fed into the Transformer. For computational simplification, large convolutions are separated in the patchify stem, and axial-attentions is used in the Transformer encoder. Figure 4.14 provides a diagrammatic overview of the proposed architecture.

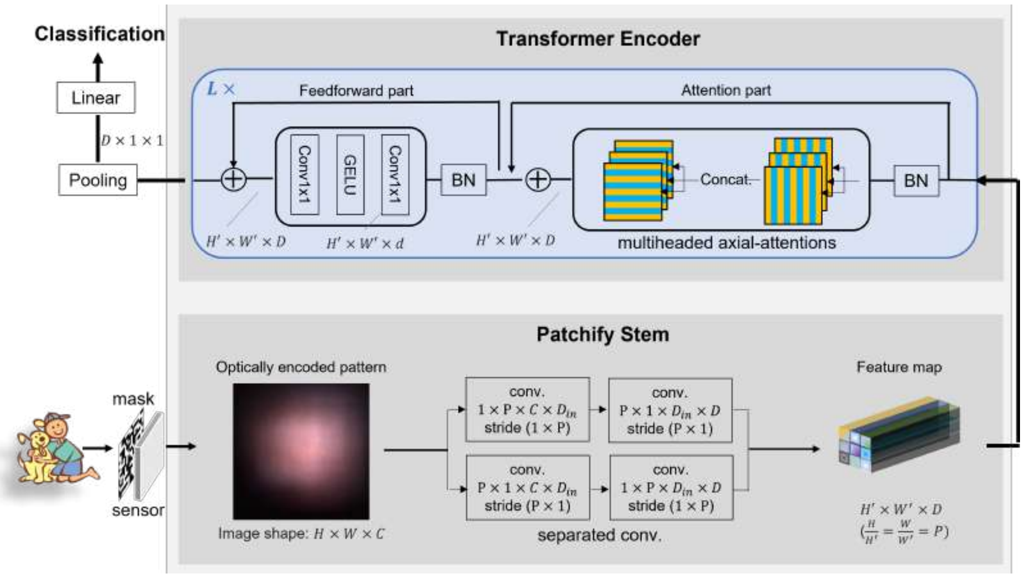


Figure 4.14: Overview of LRT

The patchify stem is first applied to the input image $\mathbf{x} \in R^{H \times W \times C}$ to obtain a feature map $\mathbf{x}' \in R^{H' \times W' \times D}$ with reduced size, where C, D are channel numbers, H, H' are height and W, W' are weights. 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 the feature map, the kernel-$(P \times P)$ is typically larger than conventional small convolution used in FCNs. Large convolution makes the training process volatile [69]. To avoid this issue, the large convolution $conv^{D}_{P \times P \times C}$ with shape $P \times P \times C$ and number D is decomposed into separated convolutions $conv_{sep}$. Applying $conv_{sep}$ to the input image $\mathbf{x}$ is expressed as follows [70]:

$$\begin{aligned} \mathbf{x}' &= conv_{sep}(\mathbf{x}) \\ &= conv^{D}_{P \times 1 \times D_{in}}(conv^{D_{in}}_{1 \times P \times C}(\mathbf{x})) + conv^{D}_{1 \times P \times D_{in}}(conv^{D_{in}}_{P \times 1 \times C}(\mathbf{x})), \end{aligned} \tag{4.2}$$

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 identical to the one used in the LIT (Section 3.3). The Transformer is followed by an average pooling layer and a linear projection:

$$result = linear(pool(\mathbf{z})), \tag{4.3}$$

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

4.3.2 Simulated encoded pattern dataset generation

Lacking inductive biases inherent to FCNs, the Transformer requires more training data. Though this issue is partially alleviated by introducing the patchify stem and MHAA in the LRT, the amount of required training data is 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 optimally initialize the weights which contributes to a higher learning efficiency in the target dataset with a relatively small scale. Accordingly, pretraining is desired in our case. However, I find that for encoded pattern recognition, pretraining on normal image datasets does not work because the encoded pattern and normal image differ greatly in the regularity of pixel arrangement. Creating encoded pattern datasets for pretraining is demanding. Nonetheless, unlike normal images which can be pulled off the internet, there is no a ready-to-use encoded pattern dataset. I propose to generate simulated encoded pattern dataset from an available normal image dataset by applying the forward model of the mask-based lensless optics.

By assuming that the PSF is shift-invariant and all incoming light is able to cast a complete pattern on the sensor, Eq. 2.1 can be simplified as

$$\mathbf{x} = \mathbf{o} * \mathbf{a}. \tag{4.4}$$

The encoded pattern $\mathbf{x}$ is approximated as the convolution between object $\mathbf{o}$ and PSF $\mathbf{a}$. To generate a simulated encoded pattern, a normal 2D image is used as the object, and the PSF is captured by illuminating the lensless camera with a point light source. For computational efficiency, Eq. 4.4 is implemented in the frequency domain:

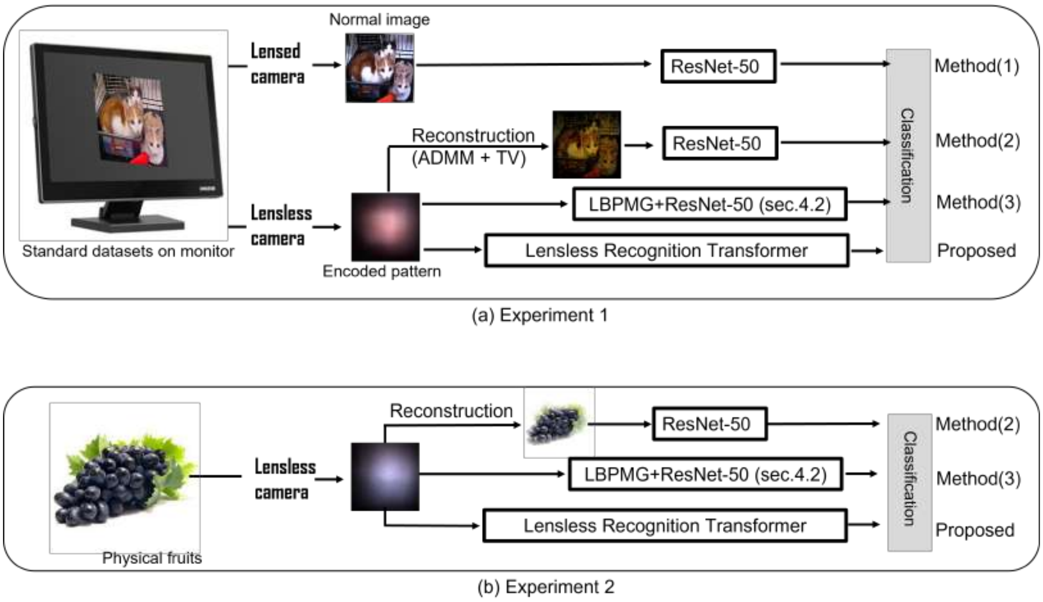
$$\mathbf{x} = \mathcal{F}^{-1}(\mathcal{F}\mathbf{o}\,\mathcal{F}\mathbf{a}), \tag{4.5}$$

with $\mathcal{F}$ and $\mathcal{F}^{-1}$ being a Fourier transform and inverse Fourier transform, respectively. Before performing the Fourier transform, zero-padding is applied to both the normal image and PSF to keep them the same size. The zero-padding area is set sufficiently large enough to satisfy the implicitly assumption of periodic behavior in the Fourier method [22].

Though the generated simulated encoded pattern is not identical to the realistic one, the pixel arrangement regularity 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 a relatively small scale.

4.3.3 Experiments

Two optical experiments are conducted, as illustrated in Figure 4.15. The used lensless camera is identical to the one introduced in Section 3.4.1. The first exper-



(a) Experiment 1

(b) Experiment 2

Figure 4.15: Diagrammatic overview of two experiments

iment compares methods on standard datasets. The image is displayed one-by-one on a monitor in front of the camera. Method 1 uses the lensed camera. The result of Method 1 is regarded as the baseline of object recognition. Method 2 is the conventional manner of applying a lensless camera to objection recognition, which reconstructs image before inference. The result of Method 2 is regarded as the baseline of lensless-camera-based methods. ResNet-50 [36] is used as the classifier in Method 1 and 2. Method 3 is the "LBPMG + FCN", as introduced in Section 4.2. In Method 3, ResNet-50 is also selected as the classifier. For the LRT, the details are listed in Table 4.5.

Another experiment evaluates the feasibility of the lensless camera (Method 2, Method 3 and the LRT) on physical objects. I built a fruits dataset by collecting 8 fruit classes from the ILSVRC-2012 ImageNet. I first train models as the first experiment. Then physical fruits are used for test.

Table 4.5: Details of implemented LRT

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

4.3.3.1 Pretraining

Pretraining is applied for all models in both Experiment 1 and 2. ResNet-50 used in Method 1, 2 and 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 Experiment 1 and 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 pixels and zero-padded to 448×448 pixels. In each epoch, the captured PSF, whose original size is 1,600×1,600 pixels, is resized to a size ranging from 150×150 pixels to 224×224 pixels before zero-padded to 448×448 pixels. With Eq. 4.5, a 448×448 pixels encoded pattern is calculated. The center 224×224 pixels 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 the first column of Figure 4.16.

4.3.3.2 Experiment 1

Experiment 1 compares the LRT with three other methods on standard datasets. The used datasets are Fashion MNIST [71] and cats-vs-dogs dataset [72]. Fashion

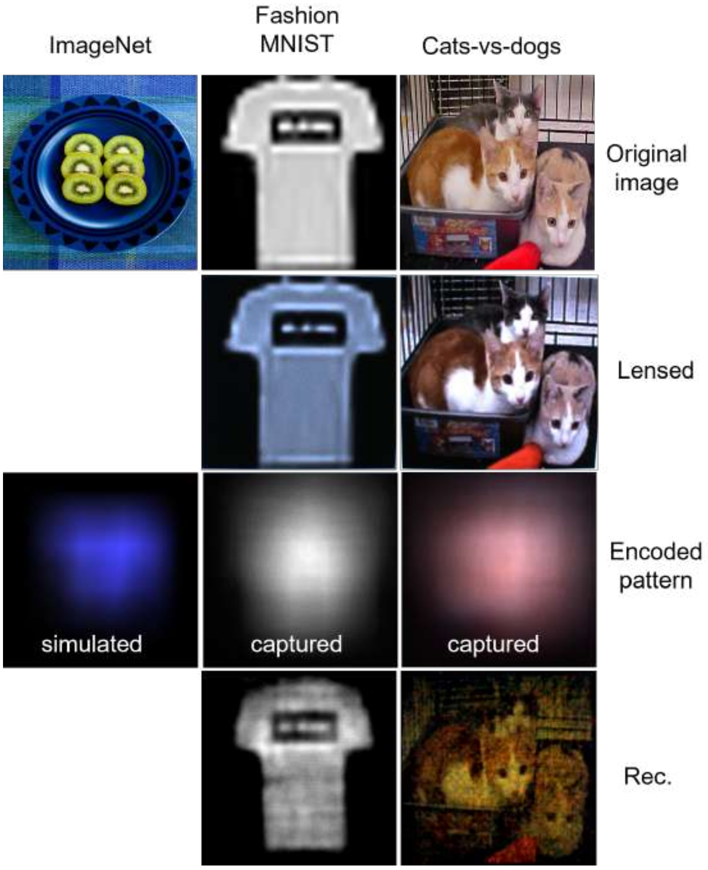


Figure 4.16: Examples of used images

MNIST dataset consists of a training set of 60,000 examples and test set of 10,000 examples. Each example is a 28×28 pixels grayscale image, associated with a label from 10 classes. The cats-vs-dogs dataset consists of 25,000 labeled colorful photos: 12,500 cats and 12,500 dogs. The 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 backgrounds without cropping or aligning.

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

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

For the lensless camera, the captured PSF and encoded pattern have valid area of 1,600×1,600 pixels. This area is cropped and resized to 224×224 pixels. Method 3 and the proposed method use the encoded pattern directly while Method 2 reconstructs images first. In Method 2, PSF-based image reconstruction with ADMM and total-variation regularization is employed. The reconstructed image has a size of 224×224 pixels, identical to the size of used PSF and encoded pattern. The reconstruction convergence chart, where each point calculates the MSE between reconstructed images in current iteration and previous iteration, is shown in Figure 4.17. Reconstruction with 10 iterations is selected for use. An example of reconstructed image is illustrated in the last row of Figure 4.16.

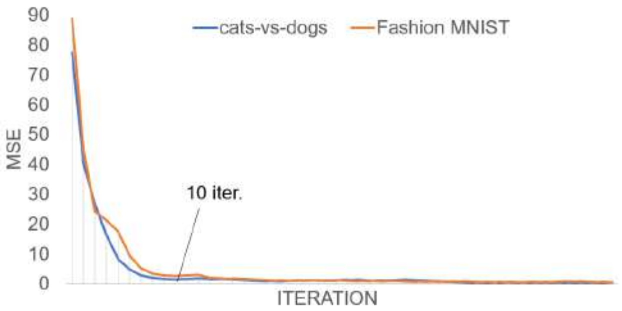


Figure 4.17: Reconstruction convergence chart

4.3.3.3 Experiment 2

Experiment 2 evaluates the feasibility of the lensless methods on physical objects. A fruits dataset is constructed 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. The models are first trained on the fruits data with the same experimental configuration as the Experiment 1. Then, test on physical fruits is conducted. I prepare two 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 to 40 cm based on different fruit sizes.

4.3.3.4 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 and Pytorch 1.7.1. ResNet-50 is trained by the Adam optimizer with β_1=0.9, β_2=0.999, a weight decay of 0.1, and a mini-batch size of 64. The proposed LRT is trained by the 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 4.6 where used epochs for convergence and runtime for one epoch are also listed.

Table 4.6: Training process comparison

Task	Method	Epochs	Time/Epoch
Clothing classification	Met. 1 (Lensed)	15	1 min
	Met. 2 (Rec.)	15	1 min
	Met. 3 (LBPMG + FCN)	25	1 min
	LRT	20	4 min 30 s
Cats-vs-dogs classification	Met. 1	15	1 min
	Met. 2	15	1 min
	Met. 3	45	1 min
	LRT	15	4 min 30 s

4.3.3.5 Results

Table 4.7: Results of Experiment 1

		Acc./AUC (%)	Time (s) for 10 images		
			Rec.	Infer.	Total
Clothing classification	Met. 1	91.50 / -	-	0.99	0.99
	Met. 2	91.12 / -	1.46	0.99	2.45
	Met. 3	90.01 / -	-	0.99	0.99
	LRT	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
	LRT	94.26 / 96.64	-	1.04	1.04

The result of Experiment 1 is listed in Table 4.7. Both accuracy/ ROC AUC (area under the receiver operating characteristic curve) and runtime are compared. All computations, including reconstruction and inference, are timed on an Intel Xeon E5-2698 v4 CPU (2.2 GHz). The proposed LRT presents a higher prediction accuracy and shorter runtime than the reconstruction-including lensless camera (Method 2). It verifies that image reconstruction is not intrinsically needed in terms of inference. Bypassing reconstruction not only saves computation but also benefits recognition performance. Method 3 is theoretically designed for a cropped and aligned target without a complex background. It evidently works well on Fashion MNIST where the target is aligned on a black background without noise, but fails on cats-vs-dogs dataset where the target is in a complex real-world scene. The results of the LRT is marginally worse than that of the lensed-camera-used method (Method 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 that of Method 1 if the optical hardware or classifier is further optimized. For example, suppressing diffraction or optimizing mask design can result in better optical signal sampling, thereby enhancing recognition. If computational resources permit, smaller patch size and deeper network could contribute to higher predictive accuracy.

Scenes, captured encoded patterns and reconstructed images of Experiment 2 are shown in Figure 4.18. The result is illustrated in Table 4.8, where top three class probability predictions for each item are listed. The LRT achieves 13/16 accuracy. The confidence is not high for some correct predictions owing to domain difference between training data (2D images displayed on the monitor) and test data (physical 3D objects). The result validates that the proposed method can generalize to physical objects. However, for robust prediction in practical application, training with physical objects is needed. For both Method 2 and 3, only 3 out of 16 predictions are correct. Method 3 has poor performance because it is theoretically inapplicable for the non-aligned scene as discussed before. For Method 2, the result is much worse than that in Experiment 1. Besides the domain difference, the main reason is that the image reconstructed by the optimization method has a much lower quality in a real-world scene. In real-world scenes where the light conditions is complex, the basic assumption of shift-invariant PSF is significantly difficult to obtain as high-angle incident light is inevitable if the position of the target is not carefully selected.

4.4 Summary

If the ultimate goal is a computer vision task such as object recognition, image reconstruction is not necessarily needed. Bypassing reconstruction benefits runtime and predictive accuracy because reconstruction not only requires additional computation but also introduces errors and artifacts. This chapter addresses reconstruction-free object recognition solutions for the lensless camera for the first time. Two solutions are proposed. The first solution is developing the computationally efficient LBPMG data preprocessing method (Section 4.2), which alleviates disturbance amplification in the lensless optics, for a FCN. Working with the LBPMG, the FCN achieves a higher predictive accuracy for cropped and aligned targets with flat backgrounds. However, the FCN, whose working mechanism heavily relies on the locality prior, is challenging in handling the encoded pattern where locality lacks. This mismatching leads to FCN's low efficiency in more complicated scenes even when working with the LBPMG. Therefore, I abandon FCNs and turn to design Transformer-based neural network which is free of the locality bias. The proposed LRT (Section 4.3), which successfully performs object recognition with complicated backgrounds without cropping or aligning, is considered a more advanced solution.

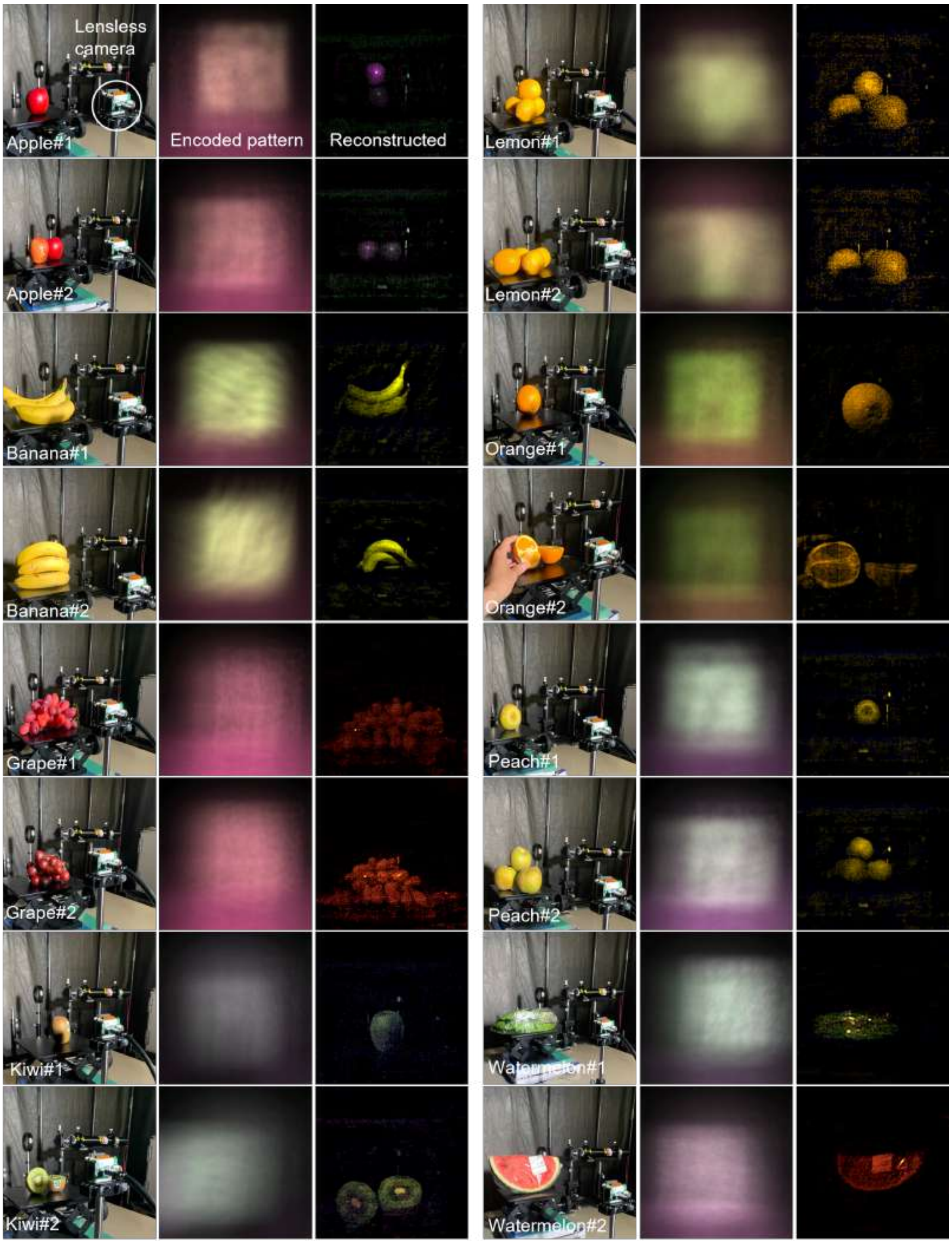


Figure 4.18: Scenes, encoded patterns and reconstructed images of Experiment 2

Table 4.8: Results of Experiment 2

Item	Met. 2 (Rec.)	Met. 3 (LBPMG)	LRT
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

5 | Mask Pattern Optimization

In the mask-based lensless camera, the mask can be any optical encoding element such as an amplitude and phase mask. The amplitude mask is a mask with distributed apertures, modulating the incident light by either passing or blocking. The phase mask is a transparent material with different heights at different locations. The mask design objective is to realize a high-performance PSF to improve the reconstruction quality or benefit a desired computer vision task. Factors including transmission rate of light, diffraction effect, signal-to-noise ratio, influence on reconstruction algorithm, and ease of fabrication are typically considered in mask design. The phase mask is more versatile in design and has higher light throughput, whereas the amplitude mask is the most commonly used because of its ease of fabrication. All masks used in this thesis are amplitude masks, and this chapter focuses on amplitude mask design.

For amplitude mask design, the aperture size can be optimized by considering the diffraction effect using the Fresnel number (Eq. 3.4). Designing the aperture distribution (or mask pattern) is a task that has attracted significant attention. Here are some notable examples of mask pattern design:

1) **Pseudorandom.** The mask with pseudorandom distributed apertures is commonly used. It is also used in all experiments in Chapter 3 and Chapter 4. Pseudorandom aperture distribution results in pseudorandom sampling. It allows image reconstruction to be compressed-sensing-based signal processing.

2) **URA and MURA.** Uniformly redundant array (URA) [11] and modified URA (MURA) [73] are designed to have a near-flat Fourier spectrum if diffraction is not considered, which provides a relatively well-conditioned

forward model. It benefits image reconstruction which is solving an inverse problem.

3) **Separable mask pattern.** This mask pattern is designed for the PSF-based reconstruction method with iterative optimization. With the separable mask pattern, the forward model can be simplified as a convolution along the rows of the image followed by a convolution along the columns [4]. In matrix form, this operation can be written as a product of a 2D image with a few small 2D matrices. Therefore, the computational complexity of image reconstruction is greatly reduced.

4) **Fresnel zone apertures (FZA).** This mask pattern is constructed to be a Fresnel zone plate. It is designed for PSF-based reconstruction method with Moiré decoding [6, 7, 26] which allows fast reconstruction. Multiplying the sensor capture with a virtual FZA results in overlapping Moiré fringes. Fast reconstruction is achieved by applying a 2D Fourier transform on the Moiré fringes.

In this chapter, I propose a novel mask pattern optimization method which optimizes the mask pattern through learning. The proposed method could be used for both image reconstruction and reconstruction-free objection recognition, provided that a DNN is applied. Section 5.1 introduces the mask pattern optimization method and Section 5.2 verifies the method through simulated experiments on a reconstruction-free objection recognition task.

5.1 Learned Optimization

Since the encoded pattern can be approximated as a convolution between the object and mask, the mask can be regarded as a convolutional filter. When a DNN is applied for the encoded pattern analysis, the mask can become a convolutional layer placed ahead of the following network. The amplitude mask pattern is a binary array representing the distribution of apertures in the mask. In this binary array, "1" represents an aperture where incoming light can pass through, and "0" represents the non-aperture area where light is blocked. The mask pattern and weights in the following networks can be optimized simultaneously through training. Note that this is not generalized optimization but a task-driven optimization for the defined task. After training, the optimized mask pattern can be printed. The aperture size optimization is only considered during printing.

As the mask pattern is required to be a binary array, its weight updating is different from that of the following networks where all weights are real-valued. Mask pattern array weights are set as real values with a constrained minimum and maximum values of 0 and 1, respectively, during backpropagation. During the forward propagation, weights are determined by:

$$\mathbf{w}^{\mathbf{b}} = \begin{cases} 1 & x \geq 0.5, \\ 0 & otherwise, \end{cases} \tag{5.1}$$

where $\mathbf{w}^{\mathbf{b}}$ is the desired mask pattern weights.

5.2 Simulated Experiments

A reconstruction-free object recognition experiment is performed in simulation for verification, shown as Figure 5.1. The task is digit recognition with MNIST dataset. As I aim at verifying the proposed mask pattern optimization method, a commonly used and simple FCN, ResNet-18 without special data preprocessing is applied as the classifier. The mask size is set as 28×28, and the initial mask pattern array is binary pseudorandom. The image from MNIST is gray and has original size of 28×28 pixels. Gaussian noise is added to create a 70 dB signal-to-noise ratio. The image is zero-padded to 56×56 pixels. I compare a fixed pseudorandom mask without optimization and the learned optimized mask. The training applies Adam optimizer with β_1=0.9, β_2=0.999 and mini-batch size of 64. After training, the optimized mask pattern is gained. Calculated on the test dataset, the accuracy of the pseudorandom mask case is 67.8% while optimized mask case achieves 84.97% accuracy.

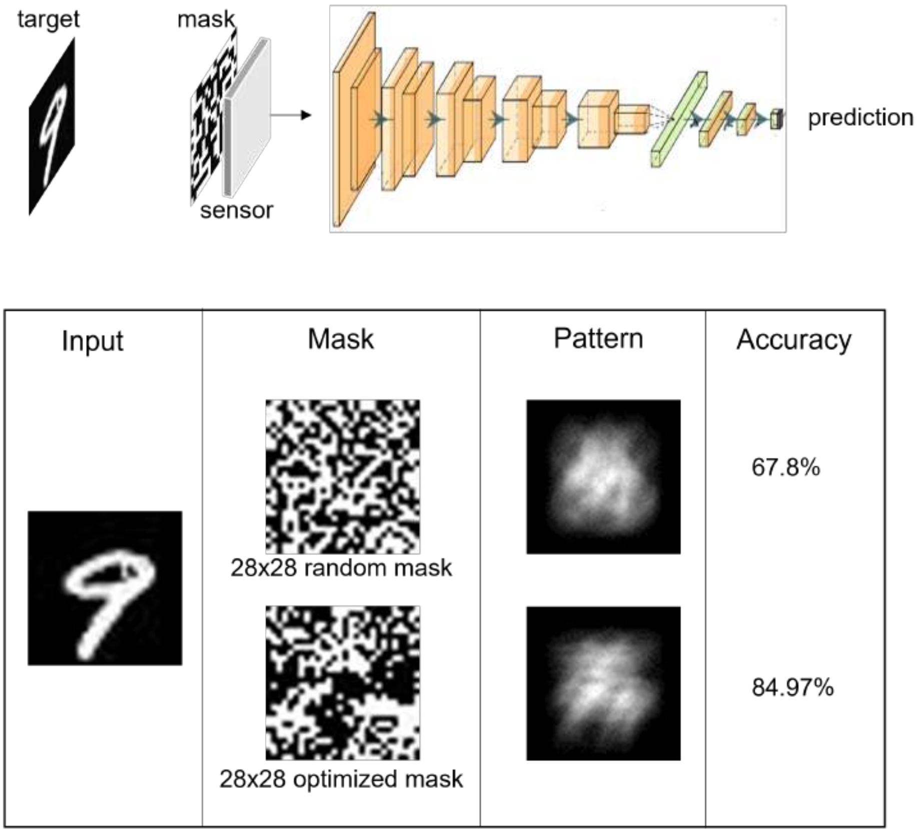


Figure 5.1: Simulation experiment and result

6 | Applications

At present, the lensless camera is unable to produce images as high-quality as the lensed camera. In photography for appreciation, the lensed camera, which has already achieved bright, sharp and aberration-free imaging, is preferred. At this time, the lensless camera aims at functional applications, such as monitoring and automation-purpose imaging and sensing. For these applications, the lensless camera is competent enough while is advantageous in terms of volume, weight and cost. The additional post-capture computation required by the lensless camera could be done through cloud computing considering that currently most of electronic devices are networked. In scenarios where the lensed camera is impractical, such as when extreme volume and weight are imposed or invisible imaging is needed, the lensless camera could be a promising or even inevitable choice.

The lensless imaging can also provide optics-level encryption. The captured encoded pattern is meaningless to human eyes. To reconstruct image with PSF-based methods, the PSF information should be known. The encoded pattern can be seen as the result of optics-level encryption by the mask. The PSF information becomes the only key to revealing the reconstructed image. When the mask pattern is a sufficiently large 2D matrix, cracking the key is difficult enough. This optical encryption is safer than the digital encryption which performs an encryption algorithm on captured images. Because for always-networked devices, there is always a risk of information leaking once digital visual data is generated. Though DNN-based reconstruction methods need no PSF information, they require a huge dataset including ground truth images and encoded patterns of the desired lensless camera. Attacking the lensless camera with a DNN-based method seems no easier than stealing the PSF information.

In this thesis, reconstruction-free object recognition is unlocked for the lensless optics. This new functionality could further extend the feature of encryption. Computer vision tasks like object recognition are demanded everywhere because they offer tremendous benefits in terms of security, convenience and efficiency. The ultra-thin, lightweight and low-cost lensless optical hardware allows lensless cameras to be installed anywhere. The lensless camera could be primarily demanded in the cheap and miniature device whose local computational resource is limited. In this case, computationally expensive computer vision tasks must rely on cloud computing. When a cloud computing server is used, the information protection should only not be against unauthorized external attackers but also abuse by authorized agents including the cloud computing provider. It seems a dilemma that we require the cloud side to perform computer vision tasks for us while we are unwilling to share our information with the agent. With a reconstruction-free recognition scheme, we can achieve both. In the reconstruction-free recognition scheme, the cloud side requests only the encoded pattern to perform computer vision tasks. Without PSF information, the cloud side only obtains the computer vision result and no other information.

Figure 6.1: Application prospects

Now we image an application case. A nearby daily item, such as glasses, a credit card or a pen, installed with lensless camera could help save lives by automatically diagnosing disorders or illness which have an effect on body movements or facial impressions and sending out distress signals. Though cloud computing provider is continuously monitoring the user with the always-networked lensless camera, the user feels comfortable and relaxed because the reconstruction-free recognition scheme is protecting the user's privacy.

7 | Conclusion and Discussion

7.1 Conclusion

Free of lens, the mask-based lensless camera is naturally advantageous in volume, weight and cost. But at present, the imaging ability of the lensless camera is limited. The imprecise modeling of the lensless system discourages the performance of the conventional model-based reconstruction methods. Data-driven deep learning could be an alternative solution because it relies more on data rather than the modeling of the target problem. Nonetheless, the overwhelming majority of deep learning algorithms are developed for the normal scene-resembling image. They are not optimal or impractical for the optically encoded pattern produced by the lensless camera. This thesis reveals the unique multiplexing property in lensless optics and works on dedicated deep learning algorithms for the lensless camera. With the proposed deep learning algorithms, the lensless camera has considerable improvement in imaging quality (Chapter 3) and additionally unlocks the reconstruction-free recognition functionality (Chapter 4). A learned mask pattern optimization method is also proposed (Chapter 5). In the end, this thesis evolves the lensless camera to be lite-yet-mighty by equipping it with deep learning. The reinforced lensless camera could find its irreplaceable application values as an extremely miniature and low-cost camera with qualified imaging ability and additional functionalities, such as invisible wavelength imaging, one-shot 3D imaging and optics-level encryption.

Essentially, this thesis is about computational imaging (image reconstruction) and computer vision (reconstruction-free recognition). Computational imaging is able to move partial imaging burden from optics to computation. Traditional lensed cameras, burdening all imaging on optics, have to rely on bulky and complex lens system to achieve bright, sharp and aberration-free imaging. Recent smartphone cameras with the application of computational imaging have simplified lens system and produce impressive high-quality photos after some post-capture processing. Lensless imaging, where almost all imaging responsibility is carried by the computation, takes the concept of computational imaging to an extreme. Computer vision is prepared for the machine to understand the visual world. The machine does not necessarily understand the world with a focused scene-resembling image, in the same manner as humans. With this insight, this thesis proposes the machine-efficient reconstruction-free recognition scheme where object recognition is performed directly on the encoded pattern and avoids errors and artifacts brought by the reconstruction itself.

7.2 Discussion

The mask-based lensless camera captures sufficient visual information of the scene, nevertheless, in a different format from the lensed camera. With the future advancements in computational imaging and computer vision, the lensless camera could possibly reach the same imaging and sensing performance as the lensed camera. However, currently there are two challenges in the proposed lensless camera. The first concerns optics. In incoherent imaging, the lensless camera inherently has a lower signal-to-noise rate mainly due to lower light throughput compared with the lensed camera. This weakness could inevitably lead to the gap of upper limit of imaging and sensing between the lensless and lensed cameras in dark environment. In future work, light throughput should be considered as an critical factor in mask pattern design. The other concerns deep learning. The used pure data-driven deep learning algorithms rely more on data than modeling. It provides nonsubstitutable benefits in lensless imaging and sensing but also induces constraints with the training data. In application, when the data deviates too much from the training data domain, deep learning could fail. This is a general issue in deep learning. Solutions could include expanding the data domain with more diversified training data and generalizability improvement with better framework design.

Bibliography

[1] David G Stork and Patrick 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):201–208, 2014.

[2] Michael J. DeWeert and Brian P. Farm. Lensless coded aperture imaging with separable doubly Toeplitz masks. *Optical Engineering*, 9109(May):91090Q, 2014.

[3] Nick Antipa, Grace Kuo, Reinhard Heckel, Ben Mildenhall, Emrah Bostan, Ren Ng, and Laura Waller. Diffusercam: lensless single-exposure 3d imaging. *Optica*, 5(1):1–9, 2018.

[4] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2016.

[5] Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. Phlatcam: Designed phase-mask based thin lensless camera. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1618–1629, 2020.

[6] Takeshi Shimano, Yusuke Nakamura, Kazuyuki Tajima, Mayu Sao, and Taku Hoshizawa. Lensless light-field imaging with Fresnel zone aperture: quasi-coherent coding. *Applied Optics*, 57(11):2841–2850, 2018.

[7] Tomoya Nakamura, Takuto Watanabe, Shunsuke Igarashi, Xiao Chen, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yam-

aguchi. Superresolved image reconstruction in fza lensless camera by color-channel synthesis. *Optics Express*, 28(26):39137–39155, 2020.

[8] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller. Learned reconstructions for practical mask-based lensless imaging. *Optics express*, 27(20):28075–28090, 2019.

[9] Salman S Khan, VR Adarsh, Vivek Boominathan, Jasper Tan, Ashok Veeraraghavan, and Kaushik Mitra. Towards photorealistic reconstruction of highly multiplexed lensless images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7860–7869, 2019.

[10] RH Dicke. Scatter-hole cameras for x-rays and gamma rays. *The astrophysical journal*, 153:L101, 1968.

[11] Edward E Fenimore and Thomas M Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[12] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[13] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[14] Adrian Stern. *Optical compressive imaging*. CRC Press, 2016.

[15] Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. Image reconstruction with transformer for mask-based lensless imaging. *Optics Letters*, 47(7):1843–1846, 2022.

[16] Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi. News report by wired, nikkei, tokyo tech, et al. https://wired.jp/article/mask-based-lensless-imaging/, https:

//www.nikkei.com/article/DGXZQOUC12CUO0S2A510C2000000, https://www.titech.ac.jp/news/2022/063968.

[17] Xiuxi Pan. Lensless imaging transformer repository. https://github.com/BobPXX/Lensless_Imaging_Transformer. Accessed: 2021-12-17.

[18] Xiuxi Pan, Tomoya Nakamura, Xiao Chen, and Masahiro Yamaguchi. Lensless inference camera: incoherent object recognition through a thin mask with lbp map generation. *Optics Express*, 29(7):9758–9771, 2021.

[19] Xiuxi Pan, Xiao Chen, Tomoya Nakamura, and Masahiro Yamaguchi. Incoherent reconstruction-free object recognition with mask-based lensless optics and the transformer. *Optics Express*, 29(23):37962–37978, 2021.

[20] Xiuxi Pan. Lensless inference transformer repository. https://github.com/BobPXX/LLI_Transformer. Accessed: 2021-10-01.

[21] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[22] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[23] José M Bioucas-Dias and Mário AT Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, 2007.

[24] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[25] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[26] Xiao Chen, Tomoya Nakamura, Xiuxi Pan, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yamaguchi. Resolution improvement in fza lensless camera by synthesizing images captured with different mask-sensor distances. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2808–2812. IEEE, 2021.

[27] Daniel Malacara. *Optical shop testing*, volume 59. John Wiley & Sons, 2007.

[28] Yunzhe Li, Yujia Xue, and Lei Tian. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media. *Optica*, 5(10):1181–1190, 2018.

[29] Shuai Li, Mo Deng, Justin Lee, Ayan Sinha, and George Barbastathis. Imaging through glass diffusers using densely connected convolutional networks. *Optica*, 5(7):803–813, 2018.

[30] Ryoichi Horisaki, Yuka Okamoto, and Jun Tanida. Deeply coded aperture for lensless imaging. *Optics Letters*, 45(11):3131–3134, 2020.

[31] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[32] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medi-*

cal image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[40] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021.

[41] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.

[42] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[43] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268,

1992.

[44] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Ganghun Kim, Kyle Isaacson, Rachael Palmer, and Rajesh Menon. Lensless photography with only an image sensor. *Applied optics*, 56(23):6450–6456, 2017.

[48] Allard P Mosk, Ad Lagendijk, Geoffroy Lerosey, and Mathias Fink. Controlling waves in space and time for imaging and focusing in complex media. *Nature photonics*, 6(5):283–292, 2012.

[49] Jasper Tan, Li Niu, Jesse K Adams, Vivek Boominathan, Jacob T Robinson, Richard G Baraniuk, and Ashok Veeraraghavan. Face detection and verification using lensless cameras. *IEEE Transactions on Computational Imaging*, 5(2):180–194, 2018.

[50] Bahram Javidi, Artur Carnicer, Masahiro Yamaguchi, Takanori Nomura, Elisabet Pérez-Cabré, María S Millán, Naveen K Nishchal, Roberto Torroba, John Fredy Barrera, Wenqi He, Xiang Peng, Adrian Stern, Yair Rivenson, A Alfalou, C Brosseau, Changliang Guo, John T Sheridan, Guohai Situ, Makoto Naruse, Tsutomu Matsumoto, Ignasi Juvells, Enrique Tajahuerce, Jesús Lancis, Wen Chen, Xudong Chen, Pepijn W H Pinkse, Allard P Mosk, and Adam Markman. Roadmap on optical security. *Journal of Optics*, 18(8):083001, 2016.

[51] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. All-optical machine learning using diffractive deep neural networks. *Science*, 361(6406):1004–1008, 2018.

[52] Zeev Zalevsky, Yevgeny Beiderman, Israel Margalit, Shimshon Gingold, Mina Teicher, Vicente Mico, and Javier Garcia. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Optics express*, 17(24):21566–21580, 2009.

[53] Bahram Javidi, Siddharth Rawat, Satoru Komatsu, and Adam Markman. Cell identification using single beam lensless imaging with pseudo-random phase encoding. *Optics letters*, 41(15):3663–3666, 2016.

[54] Artur Zdunek, Anna Adamiak, Piotr M Pieczywek, and Andrzej Kurenda. The biospeckle method for the investigation of agricultural crops: A review. *Optics and Lasers in Engineering*, 52:276–285, 2014.

[55] Xin Lei, Liangyu He, Yixuan Tan, Ken Xingze Wang, Xinggang Wang, Yihan Du, Shanhui Fan, and Zongfu Yu. Direct object recognition without line-of-sight using optical coherence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2019.

[56] Mariko Isogawa, Ye Yuan, Matthew O'Toole, and Kris 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*, pages 7013–7022, 2020.

[57] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[58] Tadashi Okawara, Michitaka Yoshida, Hajime Nagahara, and Yasushi Yagi. Action recognition from a single coded image. In *2020 IEEE International Conference on*

Computational Photography (ICCP), pages 1–11. IEEE, 2020.

[59] Mark A Davenport, Marco F Duarte, Michael B Wakin, Jason N Laska, Dharmpal Takhar, Kevin F Kelly, and Richard G Baraniuk. The smashed filter for compressive classification and target recognition. In *Computational Imaging V*, volume 6498, page 64980H. International Society for Optics and Photonics, 2007.

[60] Suhas Lohit, Kuldeep Kulkarni, Pavan Turaga, Jian Wang, and Aswin C Sankaranarayanan. Reconstruction-free inference on compressive measurements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–24, 2015.

[61] Kuldeep Kulkarni and Pavan Turaga. Reconstruction-free action inference from compressive imagers. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):772–784, 2015.

[62] Shuming Jiao, Jun Feng, Yang Gao, Ting Lei, Zhenwei Xie, and Xiaocong Yuan. Optical machine learning with incoherent light and a single-pixel detector. *Optics letters*, 44(21):5186–5189, 2019.

[63] Zibang Zhang, Xiang Li, Shujun Zheng, Manhong Yao, Guoan Zheng, and Jingang Zhong. Image-free classification of fast-moving objects using âlearnedâ structured illumination and single-pixel detection. *Optics express*, 28(9):13269–13278, 2020.

[64] Li Wang and Dong-Chen He. Texture classification using texture spectrum. *Pattern Recognition*, 23(8):905–910, 1990.

[65] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[66] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[67] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[68] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[69] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv preprint arXiv:2107.04589*, 2021.

[70] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2017.

[71] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[72] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[73] Stephen R Gottesman and E Edward Fenimore. New family of binary arrays for coded aperture imaging. *Applied optics*, 28(20):4344–4352, 1989.

Publication List

(directly related to this thesis)

Journal Papers

1. Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi, Image reconstruction with Transformer for mask-based lensless imaging, Optics Letters, Vol. 47, Issue 7, pp. 1843-1846, Apr. 2022. (code available: https://github.com/BobPXX/Lensless_Imaging_Transformer)

2. Xiuxi Pan, Xiao Chen, Tomoya Nakamura, and Masahiro Yamaguchi, Incoherent reconstruction -free object recognition with mask-based lensless optics and Transformer, Optics Express, Vol. 29, Issue 23, pp. 37962-37978, Oct. 2021.
   (code available: https://github.com/BobPXX/LLI_Transformer)

3. Xiuxi Pan, Tomoya Nakamura, Xiao Chen, and Masahiro Yamaguchi, Lensless inference camera: incoherent object recognition through a thin mask with LBP map generation, Optics Express, Vol. 29, Issue 7, pp. 9758-9771, Mar. 2021.

International Conferences

1. Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi, Design of optically extended convolutional neural network. ODF 2022, poster, P-OTh-35, Aug. 2022. (not-reviewed)

2. Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi, Lensless imaging and recognition with Transformer-based neural networks, OPTICS & PHOTONICS international Congress 2022, oral, IP3-04, Apr. 2022. (not-reviewed)

Domestic Conferences

1. Xiuxi Pan, Xiao Chen, Tomoya Nakamura, Saori Takeyama, and Masahiro Yamaguchi, マスクを用いたレンズレス光学系のためのTransformerニューラルネットワークによル非干渉かつ再構成不要な物体認認, 日本光学会年次学術講演会 Optics & Photonics Japan 2021, 27pAS8, Oct. 2021.
   (Best Presentation Award, http://aioptics.jp/result.html)

2. Xiuxi Pan, Tomoya Nakamura, Xiao Chen, and Masahiro Yamaguchi, Design of lensless inference camera, 日本光学会年次学術講演会 Optics & Photonics Japan 2020, 日本光学会年次学術講演会 予稿集, 15aAJ3, Nov. 2020. (not-reviewed)

Invited Talk

1. Xiuxi Pan, Optically encoded pattern reconstruction and recognition with Vision Transformer, The 11th AI Optics workshop (Optical Society of Japan), online, Mar. 2022. (http://aioptics.jp/meeting/20220301.html)

News Report

1. Xiuxi Pan, Xiao Chen, Saori Takeyama, and Masahiro Yamaguchi, "Lens-less" Imaging Through Advanced Machine Learning for Next Generation Image Sensing Solutions, WIRED, Nikkei, Phys.org, EurekAlert!, Tokyo Tech News, et al., May 2022.
   (https://wired.jp/article/mask-based-lensless-imaging/,
   https://www.nikkei.com/article/DGXZQOUC12CUO0S2A510C2000000,
   https://phys.org/news/2022-04-lensless-imaging-advanced-machine-image.html,
   https://www.eurekalert.org/news-releases/951125,
   https://www.titech.ac.jp/news/2022/063968)

(others)

Journal Papers

1. Xiuxi Pan, and Shinichi Komatsu, Light field reconstruction with randomly shot photographs, Applied Optics Vol. 58, Issue 23, pp. 6414-6418, Aug. 2019.

2. Xiao Chen, Noriyuki Tagami, Hiroki Konno, Tomoya Nakamura, Saori Takeyama, Xiuxi Pan, and Masahiro Yamaguchi, Computational see-through screen camera based on a holographic waveguide device, Optics Express, Vol. 30, Issue 14, pp. 25006-25019, Jul. 2022.

3. Xiao Chen, Xiuxi Pan, Tomoya Nakamura, Saori Takeyama, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yamaguchi, A mask-sensor-distance based super-resolution reconstruction for FZA lens-less camera, Applied Optics (under review)

International Conferences

1. Xiao Chen, Tomoya Nakamura, Xiuxi Pan, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yamaguchi, Resolution improvement in FZA lensless camera by synthesizing images captured with different mask-sensor distances, 2021 IEEE International Conference on Image Processing, pp. 2808-2812, Sept. 2021.

Domestic Conferences

1. Xiao Chen, Xiuxi Pan, Tomoya Nakamura, Kazuyuki Tajima, Keita Yamaguchi, Takeshi Shimano, and Masahiro Yamaguchi, Resolution enhancement in FZA lens-less camera using images captured with different mask-sensor distances, 日本光学会年次学術講演会 Optics Photonics Japan 2020, 日本光学会年次学術講演会 予稿集, 15aAJ4, Nov. 2020. (not-reviewed)

Grants & Awards

Grants

1. Cross the Border! Tokyo Tech Pioneering Doctoral Research Program, Oct. 2021- Sep. 2022
2. Super Smart Society Leadership Doctoral Student Scholarship, Apr. 2021- Sep. 2022
3. Tokyo Tech Tsubame Doctoral Student Scholarship, Oct. 2019- Apr. 2021

Awards

1. Best Presentation Award, Digital Society Device & System Workshop 2022
2. Best Presentation Award (AI Optics), Optics & Photonics Japan 2021