BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Thesis

# IMAGE CLASSIFICATION WITH DIRECTIONAL IMAGE SENSORS

by

**HAOCHUAN HU**

B.S., Wenzhou University, 2017

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2024

Approved by

First Reader

Janusz Konrad, PhD
Professor of Electrical and Computer Engineering

Second Reader

Roberto Paiella, PhD
Professor of Electrical and Computer Engineering
Professor of Materials Science and Engineering

Third Reader

Lei Tian, PhD
Assistant Professor of Electrical and Computer Engineering
Assistant Professor of Biomedical Engineering

## Acknowledgments

I am extremely grateful to my advisor, Professor Janusz Konrad, for his constant support, clarification and guidance in the field of computer vision, laying the foundation for my further exploration in this field. I also sincerely appreciate his constructive suggestions on my experiments, as well as the significant time he spent helping me revise and complete this thesis, in spite of his busy schedule.

I also appreciate Professor Roberto Paiella for helping me understand the background of nanosensors and continuously assisting me in completing experiments and writing in this area. I am also very thankful for his work in reviewing and revising this thesis. I would also like to thank Professor Lei Tian for the advice and help that he provided throughout the year and thank him for serving as a member of my committee.

Many thanks go to my friend Jianing Liu for his help in the field of optics, especially for providing me with simulation tools of sensor transfer functions that I needed for my research. Additionally, I would also like to thank my friend Zhenghao Sun; we collaborated and completed part of the experiments together in the early stages.

Finally, I am deeply grateful to all my family members, especially my father and mother for their unwavering support and encouragement throughout my life.

Haochuan Hu

# IMAGE CLASSIFICATION WITH DIRECTIONAL IMAGE SENSORS

HAOCHUAN HU

ABSTRACT

Traditional electronic implementations of CNNs (Convolutional Neural Networks) suffer from high power consumption and limited processing speed, hindering deployment in resource-constrained scenarios. Leveraging the power of photonics, an innovative imaging device has been developed in Prof. Paiella's lab, that integrates a standard image sensor with a photonic nanostructure (metasurface). This device has a unique and asymmetric response to the angle of incident light. Combined into an array within an imaging system, it can perform optical spatial filtering analogous to that in the first convolutional layer of a typical CNN tailored to image recognition. This filtering process relates an imaged object to the output of the sensor array by a coherent transfer function (CTF) or optical transfer function (OTF), under the illumination by coherent or spatially-incoherent light, respectively. By combining this all-optical convolutional layer with a shallow digital CNN, it is expected that the complexity and power consumption can be significantly reduced compared to an all-digital CNN.

In this thesis, we propose, numerically simulate and experimentally evaluate two types of the device targeting the problem of image recognition. First, we evaluate an angle-selective device, characterized by an OTF, in combination with a 5-layer LeNet CNN (fully-digital). Replacing the first digital convolutional layer of LeNet with the OTF results in a small performance drop (0.1-0.5% reduction in accuracy), but a significant reduction in computational complexity (28.8% fewer multply-accumulate

operations). Further reducing the digital network's complexity (OTF layer followed only by pooling, activation function and one fully-connected layer) leads to a hugely-reduced computational complexity (96.0% reduction) at the cost of a slight performance loss (0.6-0.8%). We also evaluate a phase-imaging device characterized by a CTF. We simulate the imaging capabilities of this device based on experimentally-measured parameters and test it on a real cell dataset. Compared to the fully-digital LeNet, the new architecture achieves an accuracy of 96.1% (2.5% reduction compared to LeNet) for 3 classes of cells and complexity savings of up to 98.4%. Finally, we propose a joint optimization of two parameters of a numerically-simulated CTF response and of a single-layer digital network. Although performance gains compared to using a fixed CTF are small (on average, about 0.5% points improvement in accuracy), we believe this is a promising pathway for further exploring optical-digital system co-design.

Overall, our numerical experiments, performed using realistic OTF/CTF responses, project significant reductions in system complexity while retaining high accuracy. These results remain to be confirmed by measurements with full sensor arrays, which would pave the way for efficient CNN-based visual recognition hardware for mobile applications.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | | |
|---|---|---|
| ADC | ..... | Adenocarcinoma |
| BF | ..... | Bright-Field |
| BN | ..... | Batch Normalization |
| CCR | ..... | Correct Classification Rate |
| CNN | ..... | Convolutional Neural Network |
| CTF | ..... | Coherent Transfer Function |
| DNN | ..... | Digital Neural Network |
| FC | ..... | Fully-Connected |
| GPU | ..... | Graphics Processing Unit |
| MAC | ..... | Multiply-Accumulate Operation |
| OTF | ..... | Optical Transfer Function |
| QPI | ..... | Quantitative Phase Image |
| SCC | ..... | Squamous Cell Carcinoma |
| SCLC | ..... | Small Cell Lung Cancer |
| SGD | ..... | Stochastic Gradient Descent |

# Chapter 1

# Introduction

## 1.1 Motivation and Challenges

Convolutional Neural Networks (CNNs) extract intricate features from images using a series of convolutional layers. They have been widely and successfully used in image recognition and image estimation. Their architecture typically consists of multiple convolutional layers that filter the input image and subsequent intermediate feature maps. Each convolutional layer is usually followed by a non-linear activation function to allow non-linear processing, much like performed by the human brain, and by a pooling layer to reduce the intermediate feature-map dimensionality.

The advent of powerful GPU computing in the last decade and the availability of rich datasets have allowed CNNs to approach or even surpass human-level performance in various image recognition tasks (e.g., object detection (Redmon et al., 2016) and image classification (Krizhevsky et al., 2017)). At the same time, the sharp increase in the number of convolutional layers comprising a CNN has led to a dramatic rise in the number of parameters and network complexity, triggering extensive research in neural network lightweighting, with such examples as lightweight neural networks (Howard et al., 2017) and model pruning (Han et al., 2015). However, fast inference on lightweight, limited-power networks, that are essential for many embedded, mobile and edge applications, remains a significant challenge.

This challenge opens up novel opportunities for optical computing solutions, which have been experiencing resurgence of interest (Wetzstein et al., 2020). Photonics

inherently provides ultrafast processing, operating at the speed of light, along with low power consumption and massive parallelism, making it a highly-attractive alternative solution to digital convolutional layers.

## 1.2 Proposed Solution

The high demand for computing power and system storage is a significant challenge when deploying CNNs onto lightweight mobile devices. To address these challenges in visual recognition tasks, we propose a system that combines a digital CNN with a novel image sensor array (Kogos et al., 2020; Liu et al., 2022). This sensor array contains carefully-designed metasurfaces with transfer function performing band-pass filtering along different orientations (i.e., anisotropic edge enhancement), which is analogous to the filtering operations performed in the first layer of traditional CNNs trained for image recognition, as well as in the first stage of human visual cortex that extracts low-level object features (Olshausen and Field, 1996). By replacing the first convolutional layer of a CNN with this sensor array, the combined system is expected to benefit in several ways, as detailed below.

- **Reduction of power requirements:** The ever increasing number of network layers has led to huge performance gains, but the surge in computational requirements has also resulted in a proportional energy-consumption increase. Typically, pooling layers are added between convolutional layers to reduce the feature-map dimension which, in turn, reduces the subsequent computational load. Hence, the first convolutional layer typically accounts for a considerable proportion of the system complexity due to the large input size (an image). By replacing the first digital layer with fast optical filtering with similar characteristics, one can significantly reduce complexity and energy consumption, especially in shallow neural networks.

- **Reduction of memory requirements:** Parameters of a neural network (primarily, network weights) and intermediate embeddings constitute a significant portion of cache usage in a digital implementation, which is another significant challenge when deploying on embedded devices equipped only with a microcontroller. By replacing the first digital convolutional layer with an optoelectronic device, we no longer require storage of the first-layer data and thus reduce the memory requirements, particularly in shallow neural networks where raw images often occupy a substantial portion of the overall memory.

- **Optical-digital tunability:** Due to the fixed design of many optical filters, using them as a replacement for trainable digital neural-network layers inevitably incurs performance loss. However, our proposed geometrically-tunable metasurfaces permit response adjustment within a certain range, thus allowing adaptation to different tasks and varying characteristics of datasets, and potentially closing this performance gap.

## 1.3 Contributions

In this thesis, we introduce a numerical simulation and performance evaluation of our combined optical-digital image recognition system, which comprises two distinct approaches related to different imaging modalities. Details regarding the imaging system have been extensively covered in prior publications (Kogos et al., 2020; Liu et al., 2022; Liu et al., 2023), and will be discussed in Chapters 4 and 5, respectively.

The main contributions of this thesis are as follows:

- Development of two optical-filtering simulations in *Matlab*, one for a fixed Optical Transfer Function (OTF) (Liu et al., 2022) and one for a tunable Coherent Transfer Function (CTF) (Liu et al., 2023).

- Integration of the two optical-filtering simulations with various configurations of digital neural networks, implemented in *Python*, into an image-recognition system.

- Analysis of theoretical computational gains offered by the optical-digital systems compared to corresponding fully-digital systems.

- Development of an optical-digital system with tunable CTF, and joint optimization of CTF parameters and neural-network parameters to improve image-recognition accuracy.

- Performance evaluation of all proposed optical-digital systems against corresponding baseline systems on MNIST (LeCun et al., 1998; Xiao et al., 2017), CIFAR-10 (Krizhevsky, 2009), CELL (Zhang et al., 2023) datasets.

# Chapter 2

# Related Work

## 2.1 Network Structure

Various approaches have been proposed to address the issue of high computational complexity of large neural-network models. One intuitive way is to optimize the network structure within a software implementation. For example, MobileNet (Howard et al., 2017) introduced a novel depth-wise convolution structure as a replacement for traditional convolutional layers, ultimately containing only 1/32 of VGG-19 parameters (Simonyan and Zisserman, 2014), while suffering only 0.9% loss in performance. In the latest version, called MobileNet-V3 (Howard et al., 2019), apart from structure improvements a network-architecture search algorithm was employed to create a lightweight network based on target dataset's characteristics, while maintaining high performance. Also, model pruning techniques have been extensively researched to remove redundant parameters, channels, and network layers in an effort to slim down the model. For instance, Han et al. (Han et al., 2015) described a method to reduce the storage and computational requirements by learning important connections within the network. By pruning the redundant connections, they reduced the number of parameters in VGG-19 from 138 million to 10.3 million with almost no loss of accuracy.

## 2.2 Imaging Systems

On the imaging side, camera technology has evolved to capture a wealth of information beyond just traditional images. These advancements allow for the acquisition of additional scene characteristics, such as depth information, infrared data, multi-hyper-spectral data, etc., significantly increasing the richness of visual information..

Research in computational imaging has also advanced the development and application of fast inference system. Similar to the approach proposed in this thesis, such imaging systems often extract some object features optically, and follow with a digital neural network to perform recognition/classification.

In one of the very first works in this area, Chen et al. (Chen et al., 2016) used bio-inspired angle-sensitive pixels and custom CMOS diffractive image sensors as a replacement for the first digital convolutional layer. Compared to digital CNNs, they achieved accuracy within 0.1-5.6% off the baseline, with energy savings of 90% in image sensing and 90% reduction in transmission bandwidth, across MNIST, CIFAR-10/100 and PF-83 datasets. Pad et al. (Pad et al., 2020) introduced an interesting approach to optical convolution based on amplitude masks. By adding an optical convolutional layer to the digital network, they managed to increase the recognition accuracy on Extended-MNIST dataset from 95.16% to 98.29% with much lower system complexity compared to an all-digital CNNs at similar accuracy. Shi et al. (Shi et al., 2022) proposed a lensless architecture to perform optical convolution and reported accuracy of 89.95% on Digits-MNIST with 47.2% energy savings compared to a digital implementation. Zheng et al. (Zheng et al., 2024) employed metasurfaces that are characterized by a point-spread function with multiple focal spots to replace the first convolutional layer, correspondingly reducing the total floating-point operations (FLOPs) by 94% while attaining 98.6% accuracy (0.7% point loss compared to a fully-digital implementation) on Digits-MNIST dataset and 88.8% accuracy (1.4%

point loss) on Fashion-MNIST dataset. Finally, Wang et al. (Wang et al., 2023) proposed a nonlinear, multilayer optical neural-network encoder for image sensing based on an image intensifer acting as an optical-to-optical nonlinear activation function. Such non-linear optical encoder outperforms the best-case digital linear encoder on QuickDraw dataset (79% versus 74% accuracy), without relying on a separate digital electronic processor.

# Chapter 3

# Overview of the Proposed Visual Recognition System

The diagram in Figure 3·1 succinctly describes the structure of our visual recognition system designed to predict a label for a given object. In the imaging sensor array, multiple metasurfaces with different transfer functions applied on different pixels perform fast optical filtering of the object. Subsequently, the filtered images from the sensor array are fed into a multi-layer digital neural network, resulting in final class prediction of the object.

## 3.1 Imaging System

Unlike traditional cameras, which rely solely on capturing raw images, our imaging system introduces a novel approach through optical filtering. The key innovation is the integration of each pixel of a standard image sensor array with a specially designed photonic nanostructure (metasurface) that only allows detection of light incident along a small, geometrically-tunable distribution of angles, whereas light incident along all other directions is reflected. Such devices can be used as optical spatial filters with unique transfer functions, based on the notion that different spatial-frequency components of an illuminated object correspond to optical plane waves propagating from the object along different directions.

Compared to typical, purely-digital neural-network recognition systems, such metasurfaces provide pre-processing capabilities with extremely low power consumption.

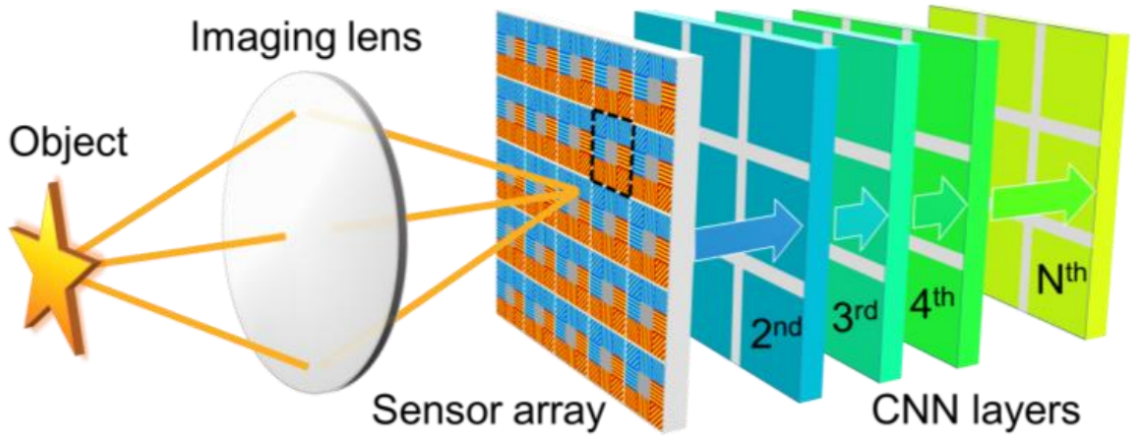


**Figure 3·1:** Proposed visual recognition hardware, where the first CNN layer is implemented with an angle-sensitive camera in the physical domain and the subsequent layers are in digital domain. The blue and orange pixels are coated with metasurfaces of type A and B, oriented along different directions.

Most importantly, the optical filtering of the metasurface can be considered an alternative to the first layer of a digital neural network. It is important to note, that a digital convolution kernel can be easily optimized for a task and dataset at hand, whereas the proposed optical filters are not as easily reconfigurable, which can result in reduced performance. Therefore, the design of the metasurface transfer function is crucial. In Chapters 4 and 5 we introduce and analyze two applications of our imaging systems for object classification, involving incoherent and coherent illumination, respectively. In the latter case, we also explore tunability of the filter transfer function.

## 3.2 Backbone Digital Neural Networks

CNNs provide a powerful framework for automatic feature learning directly from raw pixel data. Through the use of convolutional layers, batch normalization, pooling operations, and non-linear activation functions, CNNs can efficiently extract relevant features from images while preserving spatial information, with vastly fewer opera-

tions compared to a fully-connected layer. This architectural innovation has propelled CNNs to the forefront of image analysis, powering advancements in as diverse fields as medical imaging and autonomous systems.

To evaluate different optical-digital system combinations, we first implement a five-layer LeNet network (LeCun et al., 1989), which consists of two convolutional layers and three fully-connected (FC) layers, as a baseline. Then, we implement various models of extremely low complexity, for example a CNN+FC model containing only one convolutional layer, one pooling layer, one batch-normalization (BN) layer, an activation function, and one fully-connected layer. As a baseline, we feed images into either the fully-digital LeNet or CNN+FC network, and assess recognition performance of each. In the proposed system, we simulate the optical image capture and feed the resulting filtered images into a low-complexity digital network and also evaluate its recognition performance. Detailed descriptions of the proposed optical-digital systems are provided in Chapters 4 and 5.

It should be mentioned, that we also studied other reduced-complexity neural networks, including different numbers of convolution kernels, convolutional layers, fully-connected layers, etc. Taking the five-layer LeNet as the benchmark, we found that an 8-channel convolutional layer (either digital or optical) plus an FC layer can greatly reduce computational complexity and memory requirements without causing large performance loss on most datasets that we tested. At the same time, the accuracy of a CNN+FC network is much higher than that of an FC network, i.e., one using a single fully-connected layer without any convolutional layers. Therefore, in the remainder of this thesis we focus on LeNet and CNN+FC in a fully-digital form or in an optical-digital combination with two variants of directional image-sensor arrays.

## 3.3 Performance Evaluation

To analyze our designs, we adopt 4 metrics which comprehensively assess system performance, complexity, and storage requirements. The metrics related to complexity and storage apply to the simulation framework of the digital backbone implementation, but not to the optical component.

### 3.3.1 Correct Classification Rate

The Correct Classification Rate (CCR), often referred to as accuracy, measures the proportion of correctly-classified instances out of the total number of instances. It is a fundamental metric used to assess the performance of classification models, including neural networks. We will use CCR to compare performance of fully-digital and combined optical-digital image recognition systems.

### 3.3.2 Number of Parameters

The number of parameters in a neural network (also known as model size) refers to the total count of trainable weights and biases. These parameters are learned during the training process and represent the model's capacity to capture complex patterns in the data. A higher number of parameters generally indicates a larger and more complex model. The number of parameters in a convolutional layer can be computed as follows:

$$P_{Conv} = C_o \times (K^2 \times C_i + 1), \tag{3.1}$$

where $K$ is the kernel size, $C_i$ and $C_o$ are the numbers of input and output channels, and $+1$ indicates a bias added to each channel.

The number of parameters in a fully-connected layer can be calculated as follows:

$$P_{FC} = (N_i + 1) \times N_o, \tag{3.2}$$

where $N_i$ is the input size, $N_o$ is the output size and $+1$ indicates a bias added to each output of the layer.

Finally, the number of parameters in a batch normalization (BN) layer can be calculated as follows:

$$P_{BN} = 2 \times C_i, \tag{3.3}$$

where $C_i$ is the number of input channels and each channel contains one scale parameter and one bias parameter.

### 3.3.3 Multiply-Accumulate Operations

The number of Multiply-Accumulate Operations (MACs) required to compute the output of a neural network is a measure of the network's computational complexity. It serves as an estimate of the computational cost involved in running the network, which is crucial for evaluating efficiency, especially in resource-constrained environments like mobile devices or embedded systems.

The number of MACs performed by a convolutional layer can be calculated as follows:

$$MAC_{Conv} = K^2 \times C_i \times C_o \times W \times H, \tag{3.4}$$

where $W \times H$ is the size of the output feature maps.

The number of MACs performed by a fully-connected layer is:

$$MAC_{FC} = N_i \times N_o, \tag{3.5}$$

while the number of MACs performed by a batch-normalization layer is:

$$MAC_{BN} = 4 \times C_i \times W \times H. \tag{3.6}$$

### 3.3.4 Memory Usage

Memory usage in neural networks refers to the amount of memory required to store the model parameters and intermediate activations during inference. It is influenced by factors such as the network architecture, data type, and input size. Efficient memory usage is essential, particularly for deployment on devices with limited memory resources. The memory usage of a neural network can be computed as follows:

$$MEM = S \times (N_{Params} + N_{Tensors}) \tag{3.7}$$

where $S$ is the memory requirement per parameter or data sample that depends on the data type (e.g., we use the long data type which stores each parameter in 4 bytes), $N_{Params}$ is the total number of learnable parameters in a neural network, which is related to $P_{Conv}$, $P_{FC}$, $P_{BN}$, and $N_{Tensors}$ is the total size of the digital input image and of intermediate tensors at the output of each layer in the network.

## 3.4 Potential Computational Savings

In CNNs, convolutional layers are usually at the beginning of the network and are often followed by pooling layers to reduce complexity and eventually shrink the output size to the number of classes to be recognized. On the other hand, fully-connected layers typically reside towards the end of a network. Assuming that the output of a convolutional layer with complexity $MAC_{Conv}$ (3.4) is undergoing pooling by a factor of two in each dimension and then is fed into an FC layer, the input size of the FC layer is $N_i = 0.25 \times C_o \times W \times H$. Consequently, the number of MACs that must be completed by the FC layer is: $0.25 \times C_o \times W \times H \times N_o$. If the product $K^2 \times C_i$ of the convolutional layer (3.4) is greater than $0.25 \times N_o$ of the FC layer, then the number of MACs performed by the convolutional layer surpasses that of the FC layer. This is normally the case when only a few classes need to be recognized

(e.g., $N_o = 3, 7$ or 10 in our experiments), while $K^2 \times C_i$ ranges from $3^2 \times 4$ to $5^2 \times 8$. Clearly, replacing the convolutional layer by its optical variant would lead to significant savings in computational complexity of the overall system

As per equations (3.1) and (3.2), the number of parameters in a convolutional layer is typically much lower than that in subsequent fully-connected layers due to the small kernel size $K$ and number of channels $C_i, C_o$ compared to $N_i \times N_o$ (especially $N_i$ which is typically large). This results in a relatively small portion of the total number of parameters allocated to the convolutional layer, even in shallow CNNs.

Recall that according to (3.7), the memory requirement depends on the number of parameters and the size of the input image and intermediate tensors, indicating that memory savings due to the use of optical convolution include the savings due to the removal of the first digital layer (its parameters) and of the input image. However, as pointed out, the number of parameters in the first convolutional layer is a small portion of the total number of parameters, while the input image size is typically smaller than the total size of tensors produced by various layers (e.g., convolutional, batch normalization, ReLU). Hence, by replacing the first digital convolutional layer with its optical variant, the savings in the number of parameters and memory requirements are not expected to be as significant as the computational-complexity savings (number of MACs).

The first digital convolutional layer attempts to identify features in the input image, such as object edges. However, this is difficult to do when the input image is of low resolution. While using an input image of larger size (capturing higher resolution) would help extract finer features, it would also significantly increase the computational complexity (3.4) and memory requirements (3.7) of the first convolutional layer which, as discussed above, accounts for a large proportion of MACs in a shallow neural network. In contrast, optical filtering operates on continuous, high-resolution signals

(light) and can extract finer features compared to digital convolution, and can do so with much lower power consumption and higher speed. However, the drawback lies in the fixed design of an optical system; its characteristics cannot be as easily tuned as can those of a digital neural network. Also, while the impact of removing the first digital convolutional layer on the reduction of computational complexity and (to a lesser extent) memory size is significant in a shallow neural network, when the number of layers increases this impact proportionally fades. For this reason, the focus of this thesis is on shallow digital neural networks.

# Chapter 4

# Optical-Digital System Design for Incoherent Illumination

In this chapter, we consider optical-domain filtering of the input object under incoherent illumination by simulating the Optical Transfer Function of recently-developed metasurface sensors, followed by various digital neural networks to perform recognition of the imaged object.

## 4.1 Directional Image Sensor with Fixed Response

The approach is based on pixel arrays of plasmonic directional image sensors (Liu et al., 2022) designed to selectively detect light incident along a small, geometrically-tunable set of directions. The resulting imaging system can function as an optical spatial filter without any external filtering elements, leading to extreme size miniaturization. In the work reported in this chapter, the object is illuminated with incoherent (natural) light, and the sensor array is partitioned into identical blocks of 3×3 adjacent pixels. In each block, one device is uncoated, while the remaining 8 are coated with different metasurfaces. We use 2 different metasurface designs (labeled A and B) oriented along 4 different directions (0°, +45°, +90°, and –45°), as shown in Figure 4·2. Pixels of each type across the sensor array record a filtered image of the object in the field of view determined by their specific OTF. According to a fundamental theorem in Fourier optics (Goodman, 2005), under incoherent illumination the OTF of any optical spatial filter is nonzero at zero (DC) frequency. Therefore, to enable

band-pass filtering, as required for edge enhancement, the normalized photocurrent signal of each metasurface pixel is subtracted in the readout circuit from that of the uncoated pixel in the same block, so that the DC component of the image is canceled out. As a result, the camera simultaneously acquires 8 different band-pass filtered images of the object, to be fed into the subsequent layers of a digital CNN.

Since plasmonic devices of these specific designs are not yet available as fully-functional sensors, we characterize their behaviour through computer simulations. Let $I$ denote the spatial-domain representation of an image and let $T_k$ denote the frequency-domain representation of a filter with orientation $k = 1, ..., 4$ (i.e., the filter OTF). We perform filtering of image $I$ in the frequency domain as follows:

$$I'_k = \mathcal{F}^{-1}[\mathcal{F}(I) \cdot T_k], \quad k = 1, ..., 4. \tag{4.1}$$

where $I'_k$ is the output image filtered by directional filter number $k$, and $\mathcal{F}$ denotes the Fourier transform operation (implemented in software via the Discrete Fourier Transform). Figure 4·1 shows the pipeline of OTF filtering of an image that illustrates equation (4.1).

In this imaging system, we combine 8 different metasurface devices with distinct differential OTFs into one compound pixel. Figure 4·2(a) shows the two types of simulated OTF responses (type A and type B), each oriented along four different orientations in steps of 45°. An image of the digit "5" from Digits-MNIST dataset is shown in the middle, and the optically-filtered images $I'_k, k = 1, ..., 4$ obtained by applying equation (4.1) for both device types are shown in Figure 4·2(b). Clearly, our sensor effectively achieves edge enhancement for the object along different directions.

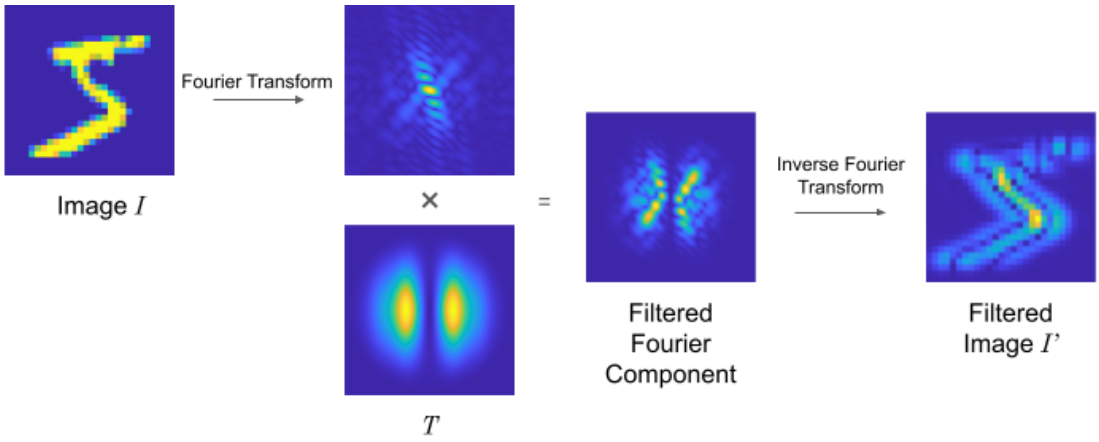


**Figure 4·1:** Computer simulation pipeline of OTF filtering described by equation (4.1). Image $I$ representing an object is Fourier-transformed and multiplied by OTF $T$ (frequency-domain filtering), and subsequently subjected to inverse Fourier transform.

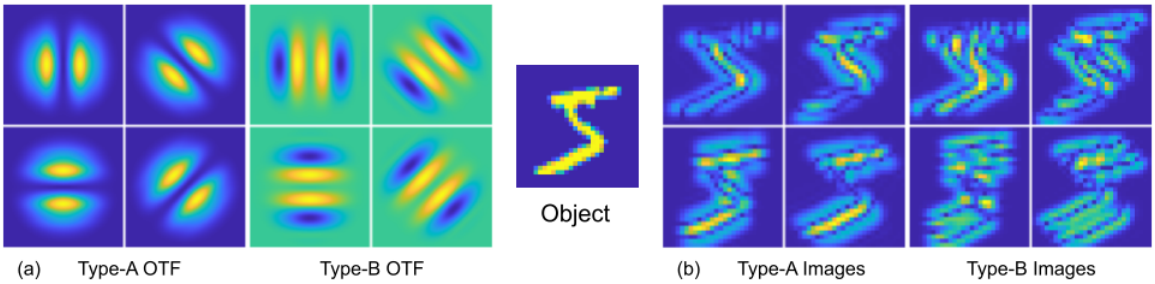


**Figure 4·2:** Different types of simulated OTFs of a novel plasmonic directional image sensor, and their impact on filtering an imaged object. (a) Differential OTFs of 2 types of metasurface devices (types A and B), each under 4 different orientations. (b) Computer-simulated output of each device type and orientation when imaging digit "5" (shown in the middle) from the Digits-MNIST dataset.

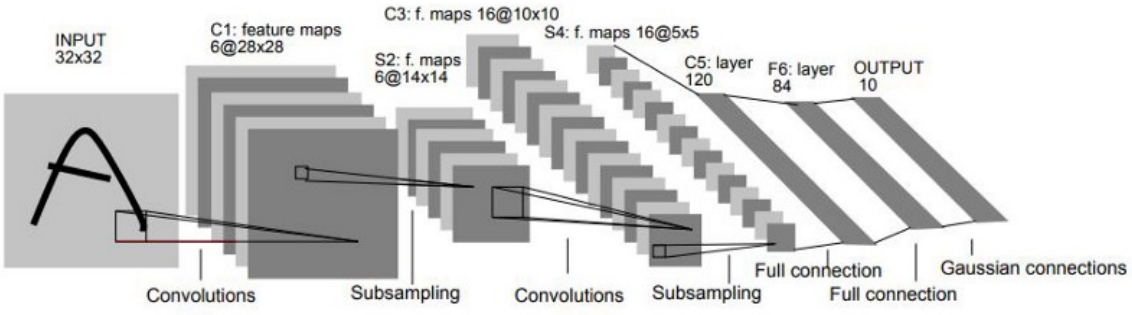


**Figure 4·3:** Architecture of LeNet-5 (LeCun et al., 1989)

## 4.2 Digital Neural Networks for Fast Inference

We use the filtered image $I'_k$ as the input to several shallow digital neural networks in order to classify the imaged object. We only consider shallow networks since the key requirement in our design is fast processing.

We study two fully-digital neural networks as a baseline for comparison with the proposed optical/digital designs:

- **LeNet** (LeCun et al., 1989): LeNet-5 is a 5-layer CNN whose original network architecture is shown in Figure 4·3. Our implementation is slightly different to allow for a fair comparison with our optical-digital systems. Since our optical layer produces 8 directionally-filtered images (channels) as a replacement for the first convolutional layer, we use 8 channels (instead of 6) in the first convolutional layer of our version of LeNet-5. Table 4.1 provides a detailed description of all layers of our variant of LeNet-5. Note that we pad the input image to 32×32 by replicating boundary pixels so that the output of the first convolutional layer has the same size as in the original LeNet-5, and therefore the remainder of the original network is unchanged.

- **CNN+FC:** This is an even shallower network than LeNet-5. It consists of only one convolutional layer with 8 channels followed by a pooling layer, batch-

**Table 4.1:** Specifications of the modified LeNet-5 used as a baseline for comparison with the OTF-LeNet- system. For an input image of 1×28×28 size, the memory requirement can be computed as follows: (14,978 + 62,606) × 4 Bytes = 310,336 Bytes = 303.06 KB.

| Layer name | Output shape | Number of parameters | Number of MACs |
|---|---|---|---|
| Input | 1×28×28 | — | — |
| Conv*[8] | 8×28×28 | 208 | 156,800 |
| MaxPool | 8×14×14 | — | — |
| BN | 8×14×14 | 16 | 6,272 |
| ReLU | 8×14×14 | — | — |
| Conv[16] | 16×10×10 | 3,216 | 320,000 |
| MaxPool | 16×5×5 | — | — |
| BN | 16×5×5 | 32 | 1,600 |
| ReLU | 16×5×5 | — | — |
| FC | 120 | 48,120 | 48,000 |
| ReLU | 120 | — | — |
| FC | 84 | 10,164 | 10,080 |
| ReLU | 84 | — | — |
| FC | 10 | 850 | 840 |
| TOTAL | 14,978 | 62,606 | 543,592 |

normalization layer, activation function, and one fully-connected layer (Figure 4.2). Detailed specifications are shown in Table 4.2. We apply the same padding of the input image as in LeNet-5.

During our exploration of optical-digital system design, we have studied the OTF filtering combined with many digital-network variants. However, in this thesis we report only the most relevant designs, combining OTF filtering with either LeNet or single fully-connected layer, as described below.

- **OTF+LeNet-:** In this design, we remove the first convolutional layer (8 channels) of the modified LeNet-5 (Table 4.1) and feed 8 filtered images $I'_k$,

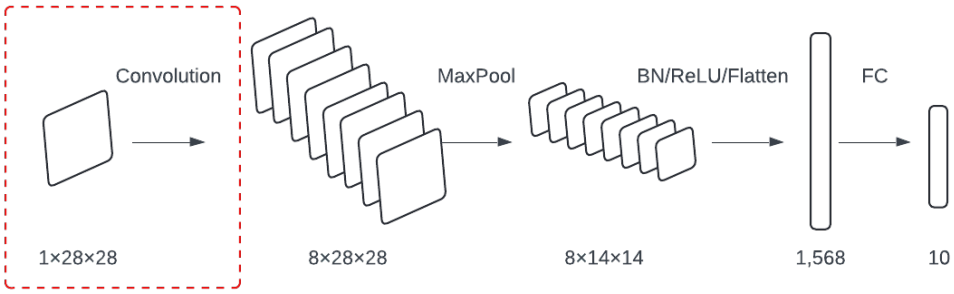


**Figure 4·4:** Architecture of a very shallow CNN+FC network, where the first digital convolutional layer (red block) can be replaced by optical convolution.

$k = 1, ..., 8$, such as those shown in Figure 4·2(b), into the second layer. Four of those images are obtained from type-A responses and four are from type-B responses. The remaining layers of this truncated LeNet design, that we call LeNet-, are identical to the modified LeNet shown in Table 4.1. Detailed specifications of OTF+LeNet- are shown in Table 4.3.

- **OTF+FC:** In this design, we input the same 8 images $I'_k$ as in OTF+LeNet-combination directly into a max-pooling layer with a 2×2 window, followed by batch normalization, ReLU, and a fully-connected layer. Detailed specifications are shown in Table 4.4.

We optimize the cross-entropy loss function to find the digital parameters (the OTF parameters are fixed and remain unchanged) using the Adam optimizer (Kingma and Ba, 2014), a variant of stochastic gradient descent, with learning rate of 0.0003 and batch size of 64. We conduct each experiment over 50 epochs, and repeat it 5 times. At each epoch, we compute the average CCR from the 5 runs, and in Tables 4.5-4.6 report the average of these average CCR values from the last 5 epochs.

**Table 4.2:** Specifications of a very shallow network CNN+FC used as another baseline for comparison with OTF-LeNet- system. For an input image of 1×28×28 size, the memory requirement can be computed as follows: (11,770 + 15,914) × 4 Bytes = 110,736 Bytes = 108.14 KB.

| Layer name | Output shape | Number of parameters | Number of MACs |
|---|---|---|---|
| Input | 1×28×28 | — | — |
| Conv*[8] | 8×28×28 | 208 | 156,800 |
| MaxPool | 8×14×14 | — | — |
| BN | 8×14×14 | 16 | 6,272 |
| ReLU | 8×14×14 | — | — |
| FC | 10 | 15,690 | 15,680 |
| TOTAL | 11,770 | 15,914 | 178,752 |

## 4.3 Datasets

We test the baseline and the proposed optical-digital designs for recognition accuracy on three datasets:

- **Digits-MNIST** (LeCun et al., 1998; Wikipedia, 2024): This is a dataset of 28×28-pixel grayscale images of hand-written 0-9 digits (10 classes). We use 60,000 training and 10,000 testing images.

- **Fashion-MNIST** (Xiao et al., 2017): This dataset also contains 28×28-pixel grayscale images of 10 types of fashion items, such as shirts, coats, trousers, snickers, sandals. Again, we use 60,000 training and 10,000 testing images.

- **CIFAR-10:** (Krizhevsky, 2009): This dataset is relatively more complex than the MNIST datasets as it consists of 32×32-pixel color images of 10 types of objects, such automobiles, airplanes, trucks, birds, dogs. It contains 50,000 training images and 10,000 testing images. In our experiments, we converted all images to grayscale for consistency among all experiments.

**Table 4.3:** Specifications of the optical-digital system OTF+LeNet-. For an input image of 1×28×28 size, the memory requirement can be computed as follows: (14,194 + 62,398) × 4 Bytes = 306,368 Bytes = 299.19 KB.

| Layer name | Output shape | Number of parameters | Number of MACs |
|---|---|---|---|
| OTF | 8×28×28 | — | — |
| MaxPool | 8×14×14 | — | — |
| BN | 8×14×14 | 16 | 6,272 |
| ReLU | 8×14×14 | — | — |
| Conv[16] | 16×10×10 | 3,216 | 320,000 |
| MaxPool | 16×5×5 | — | — |
| BN | 16×5×5 | 32 | 1,600 |
| ReLU | 16×5×5 | — | — |
| FC | 120 | 48,120 | 48,000 |
| ReLU | 120 | — | — |
| FC | 84 | 10,164 | 10,080 |
| ReLU | 84 | — | — |
| FC | 10 | 850 | 840 |
| TOTAL | 14,194 | 62,398 | 386,792 |

## 4.4 Experimental Results

In this section, we report the results of our numerical simulations for each of the fully-digital and optical-digital designs described in Section 4.2 on the three image datasets discussed above.

### 4.4.1 MNIST Results

In the first set of simulations, we used both MNIST datasets. Table 4.5 shows each design's complexity (number of parameters, number of MACs and needed memory) as well as performance expressed in terms of CCR.

On Digits-MNIST, our reference model, LeNet, achieves 99.02% CCR, whereas

**Table 4.4:** Specifications of the optical-digital system OTF+FC. For an input image of 1×28×28 size, the memory requirement can be computed as follows: (10,986 + 15,706) × 4 Bytes = 106,768 Bytes = 104.27 KB.

| Layer name | Output shape | Number of parameters | Number of MACs |
|---|---|---|---|
| OTF | 8×28×28 | — | — |
| MaxPool | 8×14×14 | — | — |
| BN | 8×14×14 | 16 | 6,272 |
| ReLU | 8×14×14 | — | — |
| FC | 10 | 15,690 | 15,680 |
| TOTAL | 10,986 | 15,706 | 21,952 |

the very shallow CNN+FC model achieves 98.53%. In contrast, our proposed optical-digital designs, OTF+LeNet- and OTF+FC, achieve respective accuracies of 99.06% and 98.26%. Similarly, on Fashion-MNIST, LeNet and CNN+FC achieve 90.22% and 89.29%, respectively, almost 10% less than on Digits-MNIST, which was to be expected since Fashion-MNIST is a more challenging dataset than Digits-MNIST. On the other hand, our proposed OTF+LeNet- and OTF+FC achieve 89.71% and 89.65%, respectively. While there is either no or slight performance loss by the new designs (up to 0.5% points in CCR), their complexity is markedly reduced. For example, OTF+LeNet- requires 27.6% fewer MACs than LeNet. Even more substantially, OTF+FC requires 87.7% fewer MACS than CNN+FC.

### 4.4.2 CIFAR-10 Results

We tested the same models on the relatively more complex CIFAR-10 dataset. While LeNet achieves a CCR of 66.22%, CNN+FC achieves only 57.06%. The optical-digital model, OTF+LeNet-, achieves 60.30% (5.9% point loss in CCR) with 27.1% fewer MACs and 1.2% less memory than LeNet. On the other hand, OTF+FC achieves only 47.98% (9.1% points loss in CCR) with 87.7% fewer MACs and 3.4% less memory

compared to CNN+FC. The complete results for CIFAR-10 are shown in Table 4.6.

**Table 4.5:** Computational complexity and performance of the fully-digital and OTF optical-digital models for the MNIST datasets.

| Model | Number of parameters | Number of MACs | Memory [KB] | $CCR$ D-MNIST | $CCR$ F-MNIST |
|---|---|---|---|---|---|
| LeNet | 62,606 | 543,592 | 303.06 | 99.02% | 90.22% |
| OTF+LeNet- | 62,398 | 386,792 | 299.19 | 99.06% | 89.71% |
| CNN+FC | 15,914 | 178,752 | 108.14 | 98.53% | 89.29% |
| OTF+FC | 15,706 | 21,952 | 104.27 | 98.26% | 89.65% |

**Table 4.6:** Computational complexity and performance of the fully-digital and OTF optical-digital models for the CIFAR-10 dataset.

| Model | Number of parameters | Number of MACs | Memory [KB] | $CCR$ CIFAR-10 |
|---|---|---|---|---|
| LeNet | 83,726 | 756,136 | 403.64 | 66.22% |
| OTF+LeNet- | 83,518 | 551,336 | 398.83 | 60.30% |
| CNN+FC | 20,714 | 233,472 | 140.95 | 57.06% |
| OTF+FC | 20,506 | 28,672 | 136.14 | 47.98% |

## 4.5 Discussion

Our optical-digital designs have achieved significant memory and complexity savings on shallow neural networks. For 28×28-pixel input images, the requirements of OTF+FC are only 12.3% in terms of number of MACs and 96.4% in memory compared to a fully-digital CNN+FC. Since the complexity (number of MACs) scales with the input image size for both CNN+FC and OTF+FC, for larger 32×32-pixel images, the number of MACs of OTF+FC remains at 12.3% compared to CNN+FC. In terms of accuracy, OTF+FC suffers only a 0.27% point loss compared to CNN+FC on Digits-MNIST but performs slightly better than CNN+FC on Fashion-MNIST. However, it suffers a 9% loss on the more complex CIFAR-10 dataset.

Our results indicate that the proposed OTF-based visual recognition system incurs minimal performance loss on simple datasets while offering large complexity reduction. However, on larger, more complex datasets it suffers a significant loss in accuracy albeit at significant savings in terms of computational complexity and required memory.

# Chapter 5

# Optical-Digital System Design for Coherent Illumination

In this chapter, we consider optical-domain filtering of transparent phase objects (cancer cells) under coherent illumination by simulating the Coherent Transfer Function of another recently-developed metasurface sensor. We also investigate the tunability of the parameters of this sensor in a narrow range that allows optimal co-design of the optical and digital components of the proposed system.

## 5.1 Compound-Eye Image Sensor with Directional Response

Another recently-developed device, a phase imaging metasensor with asymmetric angular response about normal incidence (Liu et al., 2023), has been shown to also possess directional-filtering properties similar to those of the sensors from Section 4.1. Figure 5·1(a) shows the angular response maps obtained by measuring the photocurrent signal produced by this device oriented along 4 different directions in steps of 45° as a function of the in-plane wavevector components of the incident light: horizontal - $k_x = (2\pi/\lambda)\sin\theta\cos\phi$ and vertical - $k_y = (2\pi/\lambda)\sin\theta\cos\phi$, where $\lambda = 1{,}550$ nm is the wavelength of illumination, and $\theta$ and $\phi$ are the polar and azimuthal illumination angles. The measured signal at the angle of peak detection is above 40% of the signal for identical but uncoated devices, indicating a relatively small metasurface transmission penalty. Under coherent illumination (i.e., with laser light, as appropriate for biomedical microscopy measurements of biological cells), these devices can be used

as optical spatial filters with Coherent Transfer Function (CTF) determined by their angular response map.

Similarly to the frequency-domain filtering of intensity objects described by equation (4.1), we simulate the filtering performed by each CTF as follows. Let $CTF_k$ be the Coherent Transfer Function of device number $k = 1, ..., 4$, where each value of $k$ indicates a different orientation as in Figure 5·1(a). We first crop each CTF by windowing using a circular pupil function $Q$ that describes the effect of the lens system used to image the object on the sensor array. This procedure produces the overall transfer function $T_k$:

$$T_k = CTF_k \cdot Q, \quad k = 1, ..., 4. \tag{5.1}$$

Since the numerical data $I$ contained in the CELL dataset (described in Section 5.4) correspond to phase distribution of transparent cancer cells, we first convert each object in the dataset to a phase image $e^{jI}$. This is the amplitude of the optical field propagating from each cell under illumination. Next, we perform filtering in the frequency domain, like in equation (4.1), and finally compute the magnitude squared of the filtered data (image sensors detect intensity which is proportional to the magnitude squared of the field amplitude) to produce the resulting images recorded by the sensors:

$$I'_k = |\mathcal{F}^{-1}[\mathcal{F}(e^{jI}) \cdot T_k]|^2, \quad k = 1, ..., 4. \tag{5.2}$$

## 5.2 CTF Tuning

In the experiments described in Section 4.4, the designs of the directional sensors (Section 4.1) were fixed (fixed shapes of the OTFs) - only the parameters of the digital neural networks were optimized. This optimization aimed at capturing diverse image characteristics present in the datasets used.

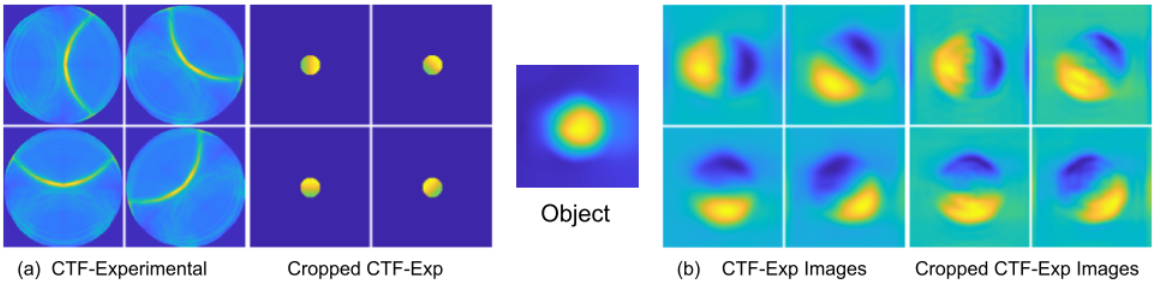


**Figure 5·1:** Experimental CTFs (full size and cropped) of a novel phase-imaging metasensor, and their impact on filtering an imaged cell. (a) Experimental CTFs of the same device oriented along 4 different directions in steps of 45°and their cropped versions. (b) Computer-simulated output of each device type when imaging a cell from the CELL dataset (shown in the middle).

The sensor described in Section 5.1 has fixed shape of the CTF as well. In both cases (OTFs and CTFs) it is unclear whether the provided transfer functions are optimal for image recognition. A tunable image sensor could enable adjustment of its transfer function for each dataset, potentially boosting performance. However, manual adjustments would require a cumbersome *adjust CTF - design DNN - adjust CTF - design DNN - ...* cycle, that would be quite impractical. A feasible approach would be to first approximate the CTF using a small number of parameters, then numerically simulate this approximate CTF and, finally, plug this simulation into the optimization loop of a digital neural network.

In order to accomplish the first step, we observe that the CTF can be analytically described by the following function:

$$R(\vec{U}) = \sqrt{B(|\vec{U}|) \times [\alpha + \beta P(\frac{|\vec{U} - \vec{S}| - n_{SPP}}{\delta})]} \tag{5.3}$$

which is a composition of a background function $B(n)$ defined as follows:

$$B(n) = \begin{cases} 1 - n^{40}, & |n| < 1 \\ 0 \quad , & \text{otherwise} \end{cases} \tag{5.4}$$

with a peak function $P(n)$:

$$P(n) = \frac{1}{1 + n^2} \tag{5.5}$$

In equation (5.3), $\vec{U}$ is the wavevector $(k_x, k_y)$ of the incident light (as defined in Section 5.1) rescaled by the constant $2\pi/\lambda$, so that its $x$ and $y$ components correspond to the normalized horizontal and vertical spatial frequencies, and $\vec{S}$ is a tunable vector that determines the peak orientation and shift from the origin. Finally, $\alpha, \beta, \delta, n_{SPP}$ are fixed parameters whose values are extrapolated from the measured response maps.

Figure 5·2 shows the simulated CTFs (both full size and cropped) with the tunable parameter $\vec{S}$ selected to reproduce the experimental CTFs of Figure 5·1, together with the corresponding filtered images of a representative object. The filtered images look extremely similar to those produced by the experimental CTFs (Figure 5·1), which suggests that the simulated CTFs should perform similarly to the experimental CTFs in image classification.

Both experimental and simulated CTFs were combined with digital layers and tested on the CELL dataset. The results reported in Tables 5.1 and 5.2 show almost identical accuracy for both CTFs. For example, when combined with LeNet and tested on CELL-3 dataset, the experimental CTF achieves 98.45% in accuracy whereas the simulated CTF produces 98.50%.

## 5.3 Joint Optimization of the CTF and Digital Layers

Thus far, we have used a fixed CTF design. However, for optimal results, the CTF needs to be adjusted to each dataset and doing so manually is complicated and time-

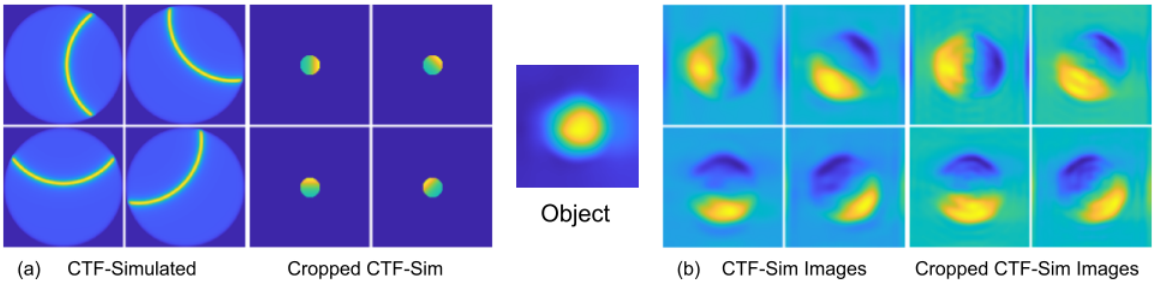


**Figure 5·2:** Simulated CTFs (full size and cropped) of a novel phase-imaging metasensor (Figure 5·1), and their impact on filtering an imaged cell. (a) Simulated CTFs based on the same design for 4 different orientations and their cropped versions. (b) Computer-simulated output of each device type when imaging a cell from the CELL dataset (shown in the middle).

consuming, as discussed above. To streamline the design of our optical-digital system, we introduce joint automatic optimization of the CTF and digital-network parameters by stochastic gradient descent. Using the filtering formula (5.2) and the CTF model (5.3), an image filtered by the simulated $CTF_k$ in channel number $k = 1, ..., N$ can be expressed as follows:

$$I'_k = |\mathcal{F}^{-1}[\mathcal{F}(e^{jI}) \cdot R(\vec{S_k}) \cdot Q]|^2 \tag{5.6}$$

where $R$ contains the tunable parameters $\vec{S_k}$.

The filtered image $I'_k$ is then sent to a simple digital neural network, consisting of a max-pooling layer with $2 \times 2$ window, a batch normalization layer, and a ReLU activation function:

$$\begin{aligned} r_k &= MaxPool(I'_k) \\ u_k &= BN_{\gamma_k,\zeta_k}(r_k) \\ c_k &= ReLU(u_k) \end{aligned} \tag{5.7}$$

where $\gamma_k$ and $\zeta_k$ are learnable scale and bias parameters of channel $k$, respectively. Finally, the ReLU outputs from all channels are sent to a fully-connected layer to

obtain the prediction:

$$\widehat{Y} = \sum_{k=1}^{n} W_k c_k + b_k, \tag{5.8}$$

where $W_k, b_k$ are learnable parameters in the FC layer and $\widehat{Y}$ is the vector of prediction probabilities for all classes.

For a more compact representation of the relationship between the input image $I$ and its class prediction $\widehat{Y}$, let $\vec{S}_{k=0,\ldots,N}$ denote tunable parameters for $N$ CTFs and let $\theta$ denote all learnable parameters of the digital layers. Then, we can combine equations (5.6-5.8) into a compact expression as follows:

$$\widehat{Y} = \text{Model}_{\vec{S}_{k=0,\ldots,N},\theta}(I) \tag{5.9}$$

By defining a loss function $\mathcal{L}_{\vec{S}_{k=0,\ldots,N},\theta}(\widehat{Y}, Y)$ between prediction $\widehat{Y}$ and one-hot-encoded ground-truth label $Y$, parameters $\vec{S}_{k=0,\ldots,N}$ and $\theta$ can be found via optimization using stochastic gradient descent.

## 5.4 Dataset

One application intended for the proposed imaging system is an extremely fast classification protocol for biological specimen. For this reason, we have selected the CELL image dataset (Zhang et al., 2023) for performance evaluation. It contains seven cancer cell lines, authenticated via the Human STR profiling cell-authentication service, including three adenocarcinoma (ADC) cell lines (H358, HCC827 and H1975), two squamous cell carcinoma (SCC) cell lines (H520 and H2170) and two small cell lung cancer (SLCL) cell lines (H526 and H69). The dataset contains 38,001 bright-field images (BF) and quantitative phase images (QPI) of cells, each of size 151×151. One sample image from each of the 7 cell lines, its Fourier transform's magnitude, as well as the results of filtering steps for our cropped experimental and simulated CTFs, are shown in Figures A·1, A·2 in Appendix A.

## 5.5 Experimental Results

### 5.5.1 CELL Results with Fixed CTFs

We evaluate the performance of all 4 network models presented in Section 4.2. In the optical-digital models, we replace the OTFs with CTFs and name these models CTF+LeNet- and CTF+FC. Note that, unlike in the case of OTFs with 8 directional responses (Figure 4·2), we have only 4 CTF orientations (Figure 5·1 and Figure 5·2). Therefore, we redesign both the first digital layer (convolutional) of LeNet- in model CTF+LeNet- and the fully-connected layer FC in model CTF+FC to accept 4-channel inputs. For a fair comparison, we also redesign the first convolutional layer of the baseline LeNet and of the CNN+FC model in the same way. In each case, we use cropped versions of either experimental CTFs (Figure 5·1) or simulated CTFs (Figure 5·2). We use the same training setup as discussed in Section 4.2.

We selected the first 2,000 QPI images from each cell line in the CELL dataset for training and the subsequent 300 QPI images for testing. Additionally, we resized the images to $144 \times 144$ by removing the first three and last four rows and columns. Since the CELL images are very smooth (and, therefore, lack high-frequency components), we decimated all images to $36 \times 36$ by averaging over $4 \times 4$ windows to reduce the computational complexity of all networks. By using each of the 7 cell categories, we created a 7-class dataset for our experiments with 14,000 training images and 2,100 testing images. We also reorganized the same images into three major categories based on 3 subtypes (ADC, SCC and SCLC) to form a 3-class dataset, with 14,000 training images (6,000 for ADC, 4,000 for SCC, 4,000 for SCLC) and 2,100 testing (900 for ADC, 600 for SCC, 600 for SCLC).

In these experiments, we use cross-entropy loss function and the Adam optimizer (Kingma and Ba, 2014) with learning rate of 0.003 and batch size of 128. We conduct each experiment over 100 epochs, and repeat it 5 times. At each epoch, we compute

the average CCR from the 5 runs, and in Tables 5.1-5.2 report the average of these average CCR values from the last 5 epochs.

Table 5.1 shows our results for the recognition of the 7 image classes. The fully-digital LeNet achieves 96.53% CCR while the proposed CTF+LeNet- achieves 96.21% CCR with experimental CTFs and 95.47% with the simulated CTFs. The optical-digital CTF+LeNet- offers a 23.3% reduction in the number of MACs and a 5.4% reduction in the required memory compared to the fully-digital LeNet. The shallower network, CNN+FC, achieves 92.82% CCR on the same data, while the proposed CTF+FC achieves 93.10% CCR with experimental CTFs and 92.09% with simulated CTFs. However, the optical-digital CTF+FC offers a 94.7% reduction in the number of MACs and a 7.6% reduction in the required memory compared to the fully-digital CNN+FC.

**Table 5.1:** Results for fully-digital and optical-digital cropped CTF models on the 7-class CELL dataset.

| Model | CTF | Number of parameters | Number of MACs | Memory [KB] | $CCR$ 7-class CELL |
|---|---|---|---|---|---|
| LeNet | — | 106,719 | 556,268 | 480.43 | 96.53% |
| CTF+LeNet- | Experimental | 106,615 | 426,668 | 454.71 | 96.21% |
| CTF+LeNet- | Simulated | 106,615 | 426,668 | 454.71 | 95.47% |
| CNN+FC | — | 9,287 | 268,272 | 76.80 | 92.82% |
| CTF+FC | Experimental | 9,087 | 14,256 | 70.96 | 93.10% |
| CTF+FC | Simulated | 9,087 | 14,256 | 70.96 | 92.09% |

We also evaluated the same models for the 3-class image recognition (Table 5.2). The CCR obtained for all models slightly increased, but the gap between the fully-digital and optical-digital models shrank. The number of MACs and memory requirements correspondingly dropped because of the reduced number of output classes. This further increased savings in terms of complexity, e.g., for CTF+FC to 96.6% reduction of the number of MACs and 10.3% reduction of the required memory compared

to CNN+FC.

**Table 5.2:** Results for fully-digital and optical-digital cropped CTF models on the 3-class CELL dataset.

| Model | CTF | Number of parameters | Number of MACs | Memory [MB] | $CCR$ 3-class CELL |
|---|---|---|---|---|---|
| LeNet | — | 106,379 | 555,932 | 479.09 | 98.59% |
| CTF+LeNet- | Experimental | 106,275 | 426,332 | 453.37 | 98.45% |
| CTF+LeNet- | Simulated | 106,275 | 426,332 | 453.37 | 98.50% |
| CNN+FC | — | 4,099 | 263,088 | 56.52 | 96.44% |
| CTF+FC | Experimental | 3,899 | 9,072 | 50.68 | 95.81% |
| CTF+FC | Simulated | 3,899 | 9,072 | 50.68 | 96.09% |

### 5.5.2 MNIST Results with Tunable CTFs

To validate the proposed tunable optical-digital system, where both optical and digital parameters are jointly optimized, we implemented a CTF+FC system which follows the same architecture as the OTF+FC system detailed in Table 4.4 with the exception of the number of CTF channels (1, 2, 4, or 8 depending on the tested configuration, as explained below). For simplicity, we used a pupil function with unity normalized cutoff frequency, but this parameter could be easily changed.

As discussed in Sections 5.2 and 5.3, the CTF characteristics in channel $k$ ($CTF_k$) are controlled by the vectors $\vec{S}_k$. We initialize two CTFs ($CTF_1, CTF_2$) with $\vec{S}_1 = [-1.12, 0]$ and $\vec{S}_2 = [0.80, 0.80]$. These specific values were selected based on the fact that the corresponding CTFs cover low frequencies present in the MNIST dataset and, at the same time, show good directional filtering sensitivity, respectively, horizontally and diagonally. Also, through rotation of $CTF_k$ by angles 0°, +45°, +90°, or –45°, we can obtain four CTFs using single $\vec{S}_k$. This increases the number of channels to the FC layer, thus increasing diversity of the input.

In experiments, we use cross-entropy loss function and Adam optimizer (Kingma

and Ba, 2014) with learning rate of 0.003 and batch size of 256. We conduct each experiment over 100 epochs, and repeat it 5 times. At each epoch, we compute the average CCR from the 5 runs, and in Table 5.3 report the maximum value of this average CCR across all 100 epochs.

We studied the impact of optical-layer tunability on the performance of our CTF+FC system in different scenarios. Table 5.3 shows the results of an ablation study that we performed on Digits-MNIST and Fashion-MNIST datasets. When using a single CTF prototype with tunable parameters $\vec{S}_1$, the improvement on both datasets due to optical-layer tuning is about 0.5% points in CCR. When this single prototype is rotated (4 channels), the gain from tuning is more pronounced for Digits-MNIST (0.5% points) than for Fashion-MNIST (0.13% points). When two prototypes are used with parameters $\vec{S}_1, \vec{S}_2$, again the improvement for Digits-MNIST is much more substantial (1.67% points) than for Fashion-MNIST (0.08% points). Finally, when the two prototypes are rotated (8 channels), there is a small performance drop of 0.18% points for Digits-MNIST but an increase of 0.91% points for Fashion-MNIST due to tuning. While not fully consistent, these results demonstrate that by joint optimization of optical and digital parameters additional performance gains can be achieved.

## 5.6 Discussion

Similarly to the use of OTFs, the replacement of the first layer of a shallow neural network by CTFs has shown significant complexity savings on a real-image dataset. On the more difficult 7-class cell-image dataset, our CTF+FC with experimental CTF achieves an accuracy of 93.10% and offers 0.28% improvement compared to the fully-digital CNN+FC, but with 94.7% savings in computational complexity (number of MACs) and 5.8% reduction of memory requirements. These results show promise

**Table 5.3:** Performance gains due to joint optimization of CTF and FC parameters on MNIST datasets.

| Initial $\vec{S}$ | Rotation | Number of CTFs | Tunable $\vec{S}$ | $CCR$ D-MNIST | $CCR$ F-MNIST |
|---|---|---|---|---|---|
| $\vec{S_1}$ | - | 1 | - | 90.95% | 83.42% |
| $\vec{S_1}$ | - | 1 | $\vec{S_1}$ | 91.49% | 83.93% |
| $\vec{S_1}$ | Yes | 4 | - | 96.66% | 87.80% |
| $\vec{S_1}$ | Yes | 4 | $\vec{S_1}$ | 97.16% | 87.93% |
| $\vec{S_1}, \vec{S_2}$ | - | 2 | - | 94.69% | 86.13% |
| $\vec{S_1}, \vec{S_2}$ | - | 2 | $\vec{S_1}, \vec{S_2}$ | 96.36% | 86.21% |
| $\vec{S_1}, \vec{S_2}$ | Yes | 8 | - | 97.48% | 88.73% |
| $\vec{S_1}, \vec{S_2}$ | Yes | 8 | $\vec{S_1}, \vec{S_2}$ | 97.26% | 89.64% |

for potential real-life applications of our recognition system.

Furthermore, our simulated CTF approximation achieves similar complexity reductions while maintaining accuracy (93.10% points for simulated CTF versus 92.09% for experimental CTF on CELL-7 as shown in Table 5.1) which allows joint optimization of the optical and digital parameters for different datasets.

The joint optimization we proposed showed slight improvement in accuracy on both MNIST datasets. On Digits-MNIST, a 0.50-1.67% points improvement was observed, except for a 0.22% points reduction for the 8-channel case (Table 5.3). On Fashion-MNIST, a 0.08-0.91% points gain was recorded. This suggests that jointly optimizing optical and digital parameters on different datasets can be beneficial and can improve an optical-digital system's recognition accuracy over an implementation with fixed optical system.

# Chapter 6

# Conclusions

## 6.1 Thesis Summary and Conclusions

We have introduced two innovative approaches for image classification by spatial filtering: a directional image sensor with fixed response and a phase-imaging sensor with tunable response. Both approaches have been simulated and demonstrated directional edge-enhanced imaging across various datasets, such as MNIST, CIFAR-10, and CELL, with the goal of reducing computational complexity while maintaining high classification performance levels. By replacing the initial digital convolutional layer with the proposed optical filtering in popular neural-network architectures such as LeNet-5 and a basic two-layer CNN+FC, our simulations have shown a significant reduction in complexity and slight savings in the number of parameters and memory requirements but without significant loss in performance. This highlights the potential of the proposed approach for use in low-power devices, such as drones and micro robots, where efficiency and performance are critical. Moreover, our study presents a mathematical approximation of the CTF, which allows for the joint optimization of CTF and digital neural network using stochastic gradient descent. Initial tests have shown some improvements on the Digits-MNIST and Fashion-MNIST datasets, suggesting that this is a promising direction. In combination with a suitable optimization strategy this may open up new avenues for further development and testing of deeper neural-network architectures on a variety of datasets.

## 6.2 Future Work

Introduction of other lightweight network architectures, e.g., depth-wise convolutional, could lead to more efficient, compact neural networks, maximizing performance with minimal resources. Additionally, the removal of redundant connections in the networks for datasets with distinct characteristics could further reduce complexity using pruning techniques. For example, the blank areas at image boundary that surround centrally-located objects in MNIST and CELL datasets contain little-to-no information and have little-to-no impact on inference results. Clearly, network connections to these areas could be removed to reduce computational complexity.

The joint optical-digital optimization proposed in this thesis could be another avenue for future work. So far, we have observed only a small improvement for CTF response followed by a single-layer neural network on the MNIST dataset. The effectiveness of this approach should be validated on additional datasets and network models. In this thesis, we developed a mathematical model for the CTF response, but the same approach and optimization could be applied to the OTF response.

In this research, we observed that for different initial CTF parameters, the optimal CTFs obtained by joint optimization are very close to the initial ones. One possible interpretation is that, due to the computation of CTF from its parameters that relies on very different operations (e.g, exponential) than those used in the digital layers, the gradient of CTF parameters may not be in the same range of values as that of the digital-layer parameters. Thus, the digital-network parameters may converge fast while those of the CTF may remain sub-optimal. Although one could manually adjust the learning rate of parameters in either CTF or digital portion, this would be impractical. Therefore, a detailed comparison of gradients of the optical and digital parts of the system is needed, perhaps followed by finding a way to balance the learning rates of both parts of the system.

# Appendix A

# Appendix

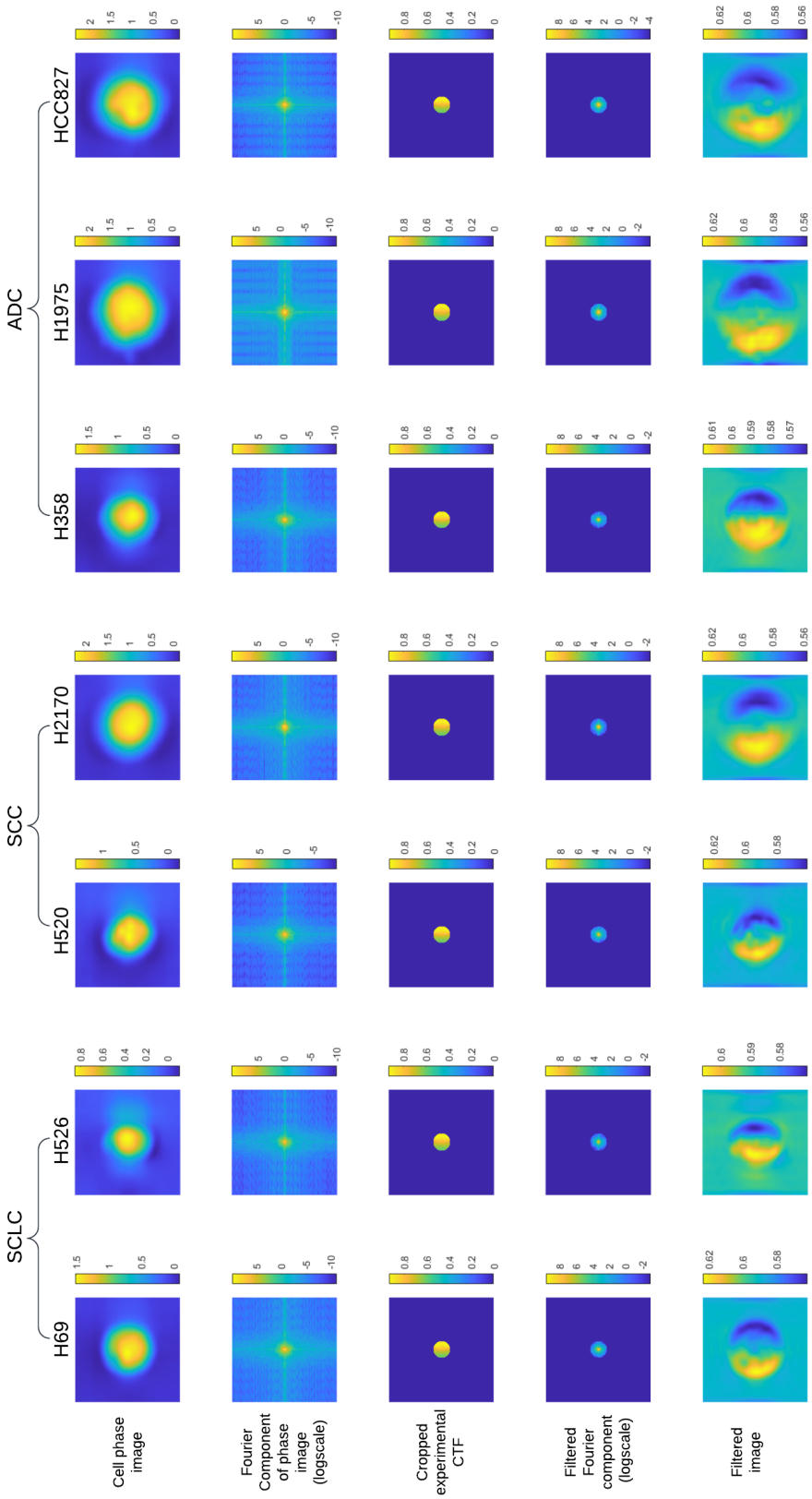


**Figure A·1:** Example of filtering steps applied to sample images from 7 classes of the CELL dataset using the cropped experimental CTF.

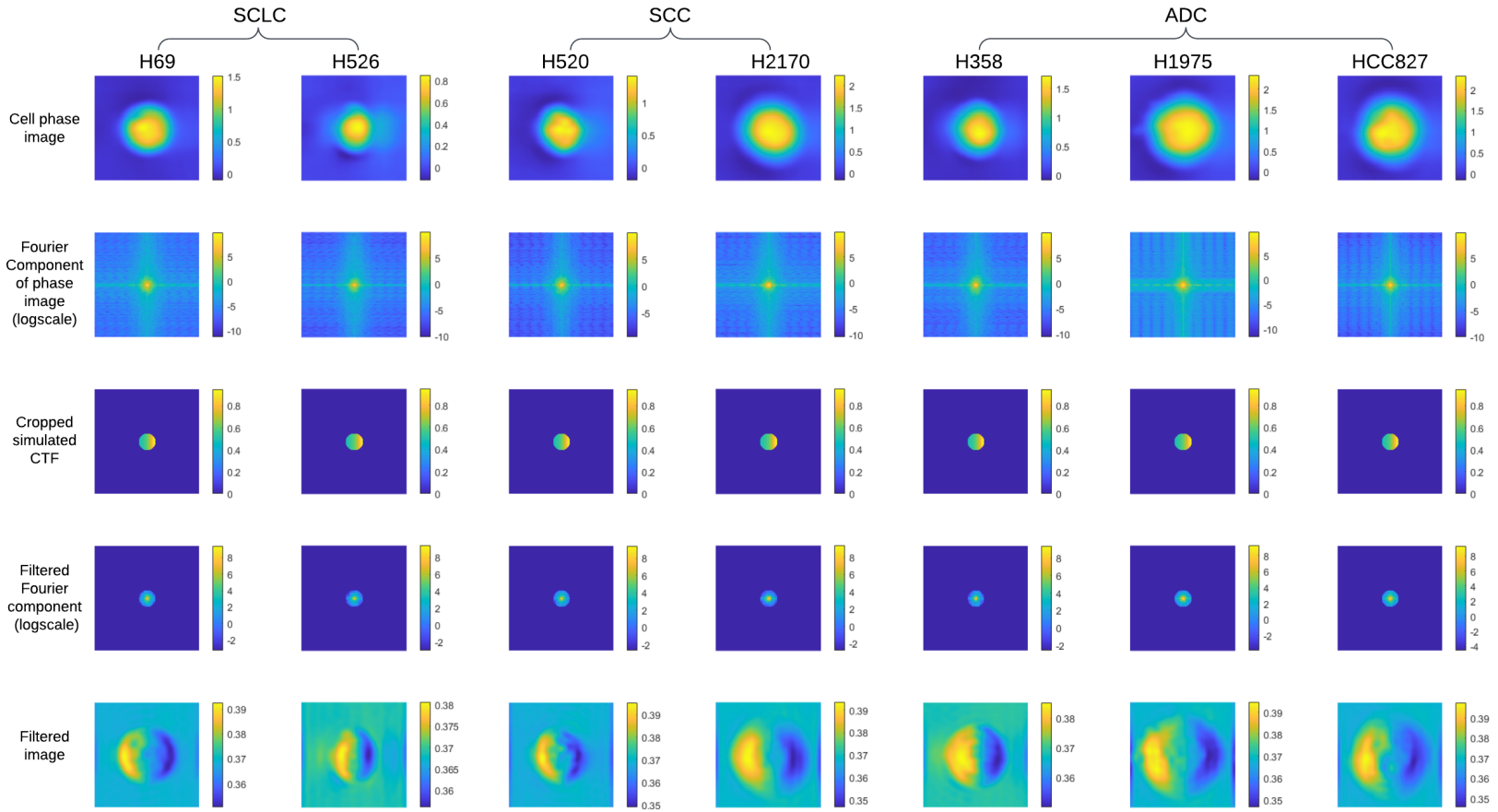


**Figure A·2:** Example of filtering steps applied to sample images from 7 classes of the CELL dataset using the cropped simulated CTF.

# References


Chen, H. G., Jayasuriya, S., Yang, J., Stephen, J., Sivaramakrishnan, S., Veeraraghavan, A., and Molnar, A. (2016). Asp vision: Optically computing the first layer of convolutional neural networks using angle sensitive pixels. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 903–912.

Goodman, J. W. (2005). Introduction to Fourier optics. Roberts and Company publishers.

Han, S., Pool, J., Tran, J., and Dally, W. (2015). Learning both weights and connections for efficient neural network. Advances in neural information processing systems, 28.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1314–1324.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kogos, L. C., Li, Y., Liu, J., Li, Y., Tian, L., and Paiella, R. (2020). Plasmonic ommatidia for lensless compound-eye vision. Nature communications, 11(1):1637.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6):84–90.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems, 2.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

Liu, J., Wang, H., Kogos, L. C., Li, Y., Li, Y., Tian, L., and Paiella, R. (2022). Optical spatial filtering with plasmonic directional image sensors. Optics Express, 30(16):29074–29087.

Liu, J., Wang, H., Li, Y., Tian, L., and Paiella, R. (2023). Asymmetric metasurface photodetectors for single-shot quantitative phase imaging. Nanophotonics, 12(17):3519–3528.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583):607–609.

Pad, P., Narduzzi, S., Kundig, C., Turetken, E., Bigdeli, S. A., and Dunbar, L. A. (2020). Efficient neural vision systems based on convolutional image acquisition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12285–12294.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788.

Shi, W., Huang, Z., Huang, H., Hu, C., Chen, M., Yang, S., and Chen, H. (2022). Loen: Lensless opto-electronic neural network empowered machine vision. Light: Science & Applications, 11(1):121.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Wang, T., Sohoni, M. M., Wright, L. G., Stein, M. M., Ma, S.-Y., Onodera, T., Anderson, M. G., and McMahon, P. L. (2023). Image sensing with multilayer nonlinear optical neural networks. Nature Photonics, 17(5):408–415.

Wetzstein, G., Ozcan, A., Gigan, S., Fan, S., Englund, D., Soljačić, M., Denz, C., Miller, D. A., and Psaltis, D. (2020). Inference in artificial intelligence with deep optics and photonics. Nature, 588(7836):39–47.

Wikipedia (2024). MNIST database — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=MNIST%20database&oldid=1216945316. [Online; accessed 08-April-2024].

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747.

Zhang, Z., Lee, K. C., Siu, D. M., Lo, M. C., Lai, Q. T., Lam, E. Y., and Tsia, K. K. (2023). Morphological profiling by high-throughput single-cell biophysical fractometry. Communications Biology, 6(1):449.

Zheng, H., Liu, Q., Kravchenko, I. I., Zhang, X., Huo, Y., and Valentine, J. G. (2024). Multichannel meta-imagers for accelerating machine vision. Nature Nanotechnology, 19(4):471–478.

# CURRICULUM VITAE

## Haochuan Hu

### Education

- M.S. in Electrical and Computer Engineering
  Boston University (BU), 2024
- B.S. in Electronic Information Engineering
  Wenzhou University (WZU), 2021

### Research and Projects

- *Image Classification with Directional Image Sensors*, BU
  Project on combined optical-digital image recognition system aiming to reduce computational complexity and power requirements compared to all-digital designs, while maintaining accuracy.
  - Implemented a computer simulation of optical filtering.
  - Evaluated performance (accuracy, computational complexity and memory requirements) of optical-digital combined system.
  - Implemented a numerical approximation of optical sensor's transfer function and developed a method for joint optimization of this transfer function and of digital neural-network parameters.
- *People Detection in Overhead Fisheye Images*, BU
  Project aiming at detecting people from an overhead viewpoint and subject to fisheye-lens distortions.
  - Implemented RAPiD, a YOLO-v3 based people-detection algorithm.
  - Improved performance of standard frame-by-frame RAPiD by enforcing temporal coherence of detections during post-processing (historical inference information).
- *Photo Enhancement Using Flash/No-Flash Image Pairs*, BU
  Project to enhance visual image quality by combining two images captured in a dark scene, one with a flash and one without.
  - Implemented joint-bilateral filter, developed a detail-preserving de-noising module and a detail-inheriting module to enhance no-flash photo by transferring details from a photo captured with flash.

- *Image Style Transfer*, WZU
  Project aiming to transfer an image style from one image to another image.
  - Built several Neural Style Transfer models for style transfer from various images based on convolutional neural networks; achieved desired transfer effects on images of various paintings.
  - Combined different networks and convolutional layers and realized integration of different style-transfer effects.
- *Micro-Expression Recognition Using Convolutional Neural Networks*, WZU
  Project targeting recognition of facial micro-expressions from image sequences.
  - Processed micro-expression datasets and constructed recognition neural-network models.
  - Modified and evaluated several classical digital neural networks, e.g., VGGNet, ResNet.
- *Travelogue and Automatic Tourist Guide App*, WZU
  Development of a smartphone application for travelogue sharing and image recognition.
  - Undertook Android App development, including function implementation, interface design and back-end server design.
  - Implemented a neural network model capable of recognizing images and then provide Eguide service based on matching keywords.