Deep Learning for Medical Imaging Applications

by

NOOSHIN MOJAB
B.S., Sharif University of Technology, 2010

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Chicago, 2021

Chicago, Illinois

Defense Committee:
Philip S. Yu, Chair and Advisor
Xinhua Zhang
Bing Liu
Joelle Hallak, Ophthalmology and Visual Science
Darvin Yi, Ophthalmology and Visual Science

This thesis is dedicated to my parents and my brothers,

for their endless love, support,

and encouragement.

ACKNOWLEDGMENTS

I would like to thank my advisor, Philip S. Yu, for his supports, encouragement, and mentoring. He provided a research environment in which I could freely explore different topics to find my passion and interest. He introduced me to great research opportunities that certainly changed my life. He has provided me with invaluable advice during my PhD years from many different aspects. I would also like to thank Joelle Hallak for her endless support and encouragement. She introduced me to the exciting world of medical research. My passion for choosing research in the medical field that can make an impact is fostered by her kindness and mentoring. Without Philip and Joelle's great help, mentoring, and friendship, I would certainly not be where I am today.

I would also like to thank my collaborators, whose help and feedback paved the way for tackle challenges in my PhD journey and become a better researcher. I would also like to thank my committee member whose comments and feedback helped me to overcome many difficulties in my research.

I would also express my deepest gratitude to my parents and brothers who endlessly supported me and encouraged me to pursue my passion. To my parents who sacrificed a lot and supported my decision to come to the US and pursue my PhD degree. To my brother who always supported me, encouraged me, and made me laugh during the most stressful time in my PhD years. Without my family's enormous sacrifices and love, I would certainly not be where I am today.

ACKNOWLEDGMENTS (Continued)

Contribution of Authors Chapters 2, 3, 4 in this work have been published previously in [1–3] and Chapter 5 is submitted to [4].

Chapter 1 outlines my dissertation research. Chapter 2 represents a published manuscript [1] for which I was the primary author and major driver of the research. Dr. Joelle Hallak and Dr. Vahid Noroozi contributed by discussing the problem and ideas with me. Dr. Joelle Hallak and Dr. Philip Yu contributed to the writing of the manuscript. Chapter 3 represents a published manuscript [2] for which I was the primary author. Dr. Vahid Noroozi, contributed by discussing the problem and ideas with me. Dr. Joelle Hallak and Dr. Philip Yu contributed to the writing of the manuscript. Chapter 4 represents a published manuscript [3] for which I was the primary author. Dr. Philip Yu and Dr. Joelle Hallak contributed to the writing of the manuscript. Chapter 5 represents a submitted manuscript [4] for which I was the primary author. Dr. Darvin Yi contributed by discussing the problem and ideas with me. Dr. Darvin Yi and Dr. Philip Yu contributed to the writing of the manuscript.

N.M

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

LIST OF ABBREVIATIONS

MRN	Medical Record Number
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
UIC	University of Illinois at Chicago
UIH	University of Illinois Hospital and Health Sciences System
OCT	Optical Coherence Tomography
PK	Primary Key
HVF	Humphrey visual field
IOL	Intraocular Lens master calculation report
DR	Diabetic Retinopathy
AMD	Age-related macular degeneration
CDR	Cup to Disc ratio
CNN	Convolutional Neural Network
ROI	Region of Interest
DSC	Dice similarity coefficient

LIST OF ABBREVIATIONS (Continued)

TR	Transfer Learning
MTL	Multi-task learning
Wide ResNet	W-ResNet
I-ODA	Illinois Ophthalmic Database Atlas
InterGD	Interpretable Glaucoma Detection
CvS	Classification via Segmentation

SUMMARY

Over the past decade, deep learning algorithms have been proved to excel at various medical imaging tasks ranging from disease detection to progression prediction and segmentation. One of the primary goals of medical imaging is to improve clinical translation enabling generalization across different real-world clinical settings and patient populations. However, there are some major challenges in real-world data applications that pull down the progress.

- Real-world medical imaging dataset: Real-world clinical data is different from standardized data collected from artificial settings, e.g. clinical trials that are commonly employed in the literature. Creating a dataset that reflects the characteristics of real-world clinical data is very challenging. The lack of real-world datasets inhibits the further improvement in clinical translation.
- Clinical interpretability: Interpretable results play an important role in decision making in clinical setting, e.g. imaging biomarkers for disease diagnosis. Therefore, providing a solution that targets the transparency of the model's solution is necessary for more reliable decision-making.
- Generalization to real-world setting: Standardized datasets collected from artificial settings do not represent real-world clinical data. This problem, raise a concern on yielding varying result in clinical translation. Moreover, lack of feature extraction networks for medical data limits the generality of learned features and hence limits the clinical translation to real-world settings.

- Learning from small datasets: In many real-world clinical applications, labels are very expensive and difficult to obtain. Therefore creating a large dataset can be very time-consuming or even infeasible, e.g. diagnosis of rare disease. However, the generalization performance of deep learning algorithms tends to degrade when trained with small datasets.

In this thesis, we explore four different settings aiming to address the above challenges in medical imaging in the context of ophthalmic imaging application: creating a real-world medical imaging dataset, multi-task learning, self-supervised learning, and classification for small datasets. In the first setting, we address the lack of real-world data by creating a real-world medical imaging dataset that is created from diverse longitudinal data from real-world setting. This dataset provides an infrastructure for validation studies and clinical translation across different clinical settings. In a multi-task learning setting, we explore clinical interpretability by formulating a real-world clinical problem in a multi-task framework. In a self-supervised setting, we develop a framework for feature extraction for ophthalmic imaging data to improve the generality of learned representations. We further explore the generalization capacity of deep learning algorithms on real-world data versus standardized datasets and analyze the importance of real-world data for clinical translation. In classification for a small dataset, we explore how to incorporate the high bias shape prior from the segmentation module into the learning process to solves the classification problem for extremely small datasets.

CHAPTER 1

INTRODUCTION

The objective of my thesis research is to develop infrastructure and decision-making systems that are capable of analyzing real-world medical imaging data and improve upon clinical translatability across clinical settings.

Medical imaging is referred to as a set of processing techniques that create visual representations of interior parts of the body such as organs or tissues. Different types of medical imaging technology give different information about the area of the body to be studied or medically treated, e.g. X-ray, MRI, retinal vessels, Fundus of the eye. Medical imaging technology has been improving significantly over the past decade producing high-quality imaging. Medical imaging analysis enhances the efficiency of clinical examination and hence leading to more accurate diagnosis and treatment which can potentially reduce invasive medical procedures. Medical imaging can help physicians with the early detection of diseases resulting in improvement of patient health and the overall human life expectancy.

Medical imaging has a very complex structure and its interpretation heavily relies on the expertise of medical specialists. However, with advancing imaging technology, the amount of data generated every day is rapidly growing and hence exceeding the scope of traditional analysis. Consequently, dependency on the knowledge of medical experts becomes less accessible and more challenging, and the decision-making process becomes more prone to human errors.

Traditional machine learning algorithms were among the first automated systems that were developed to assist medical experts in the analysis of medical imaging. However, these methods also heavily rely on domain-specific experts for feature extraction which not only increases the probability of human errors, it is also time-consuming and expensive. With computer vision advancing significantly using deep learning models and their ability to automatically extract features, we have witnessed dramatic growth in medical imaging and its applications [5–8].

One of the primary goals of medical imaging is to improve clinical translation which enables generalization across different real-world clinical settings and patient population. Although deep learning advances have helped the field of medical imaging enormously in solving many challenging problems from diagnostic detection to organ/substructure segmentation and progression predictions [9–16], clinical translation still remains a challenging problem. The difficulty of improving the clinical translation stems from several main challenges including the lack of dataset reflecting real-world clinical data, clinical interpretability, generalization capacity on real-world clinical settings, and learning from extremely small datasets that is inevitable for many applications.

The current research mainly studies the medical imaging problems in artificial settings where the data is standardized and mainly collected from the clinical trial which could limit their generalization to real-world clinical settings. However, real-world data is quite different from the standardized datasets that are commonly employed in the research studies for models development and problem solving. The real-world data is very complex, characterized by variability in quality and settings, and lacks standardization. Clinical interpretability plays a

vital role in decision-making across real-world clinical settings. However, the current research studies mainly focus on methods that predict the final result, e.g. diseased or healthy, disregarding the transparency of the model's solution to the task. The lack of interpretability in such methods makes the decision-making less reliable and limits their practical use in a clinical setting. The current work in medical imaging commonly employs the standardized dataset to develop deep learning based models for various applications. With standardized datasets not reflecting the characteristics of real-world clinical data, there has been a growing concern on the generalization capacity of deep learning models in real-world clinical settings. Although the medical imaging applications using standardized datasets have been extensively studied in the literature, there is very limited works on the generalization capacity of developed models in real-world settings. Moreover, due to difficulty of collecting labeled data in medical field, there are many applications with shortage of data that relies on pretrained networks on non-medical images such as ImageNet [17] to extract features. However, since the nature of medical data is quite different from natural images in ImageNet, the capacity of these methods for extracting effective visual representations for medical data becomes limited. In real-world clinical setting, obtaining labels is not only very expensive and difficult but also requires the domain knowledge of medical experts. Moreover, there are many applications in which the collection of a large amount of data is not only challenging but also infeasible, e.g. diagnosis of rare diseases. In such a scenario, a successful model must be able to learn from only a handful of examples. However, deep learning models show their full potential in modeling and solving problems when a large amount of data is available for training.

My thesis research mainly focuses on addressing the above challenges aiming at improving clinical translation in medical imaging, particularly in ophthalmic imaging applications. In the following sections, I will outline our work in four settings: creating a new ophthalmic imaging dataset where the data is collected from real-world clinical data and addresses the lack of real-world medical imaging datasets. Multi-task learning where we target the transparency of the model's solution to the task by achieving clinical interpretability in the context of a real-world clinical application. A self-supervised learning framework where we target the generality of learned visual representations for ophthalmic imaging data and assess the generalization capacity of deep learning models to real-world clinical settings and importance of real-world data for clinical translation. A classifier for small datasets that harnesses the power of segmentation to solve the classification problem for extremely small datasets.

1.1 Real-world Medical Imaging Dataset

The current medical imaging datasets that are most commonly used to develop deep learning based models for ophthalmic applications are collected from a standardized artificial setting. However, the real-world dataset is quite different and it is usually collected from multiple heterogeneous settings, characterized by diverse patient populations with longitudinal data, variability in quality, machine-type, setting, and source. Hence, the existing methods that are developed based on the assumption of standardized data cannot generalize well to the actual clinical data. Therefore, having a medical imaging dataset that reflects the characteristics of real-world medical data becomes the first and a crucial step to further advancement of medical computer vision and improvement in the clinical translation. In Chapter 2, we introduce a

new medical imaging dataset in the context of ophthalmic imaging applications, called I-ODA, that is created from real-world data. I-ODA is a longitudinal multi-modal dataset gathered from patient imaging data who visited the Eye and Infirmary clinic of the University of Illinois at Chicago (UIC). I-ODA aims to alleviate the shortage of real-world clinical data and provide a potential benchmark for validation study and hence improving clinical translations and advancing the state-of-the-art in medical computer vision applications.

1.2 Multi-task Learning for Clinical Interpretablity

The current research studies, in particular for diseases detection, mainly focus on predicting the final result, e.g. disease or non-diseased, disregarding the clinical interpretability of the results. This in turn limits the reliability of the decision making and hence limits their practical application in real-world clinical settings. In Chapter 3, we present a multi-task learning method to the above problem. We address the clinical interpretability in the context of an ophthalmic imaging application, glaucoma detection from an ophthalmic imaging modality, called Fundus photo of the eye. Current methods on glaucoma detection fall into two major groups, (1) classification methods that predict the presence of glaucoma directly from a Fundus photo, and (2) segmentation methods that focus on locating the imaging biomarkers identifying the disease in a Fundus photo. The methods that merely perform classification, suffer from the lack of interpretability. On the other hand, segmentation methods achieve clinical interpretability but they face the challenge of collecting datasets with segmentation labels. Segmentation labels are very expensive and laborious to obtain in the medical field. The shortage of segmented datasets limits the practical application of segmentation based approaches. In Chapter 3, we

study the problem of clinical interpretability for glaucoma detection problem. We propose a novel framework called InterGD aiming to address the two challenges, shortage of segmentation labels, and lack of interpretability by formulating the problem into a multi-task learning framework. We target the transparency of the model's solution to the task by locating two imaging biomarkers, optic disc, and cup and calculation of clinical measurement, CRD, to identify glaucoma. We also apply a masking technique to ensure the correct alignment of cup area inside the disc area. We show the effectiveness of the proposed approach in addressing the above challenges for glaucoma detection application.

1.3 Self-supervised Learning For Real-World Data Applications

Most of the existing research studies in medical imaging focus on developing deep learning based models for a broad range of applications using standardized datasets which are mainly collected from artificial settings such as a clinical trial. However, these datasets do not necessarily represent the real-world data and hence yielding varying results in clinical translation. On the other hand, the complexity of real-world data pose a major challenge for training deep learning based models. Hence, there has been a growing concern on the generalization capacity of deep learning models in real-world clinical settings. Although the medical imaging applications using standardized datasets have been extensively studied in the literature, there has been very little work on the generalization and translatable capacity of deep learning based models in real-world settings. Moreover, due to shortage of labeled data in medical field, many application with limited data formulate the problem in transfer learning setting relying on pretrained networks using non-medical images such as ImageNet for feature extraction. However, due to

dissimilarity of medical data and natural images in datasets such as ImageNet, the capacity of such methods in learning effective representations for medical data becomes limited. In Chapter 4, we aim to improve the generality of learned features for medical imaging data and assess the generalization capacity of deep learning based models on real-world data versus standardized datasets. As collecting large-scale labeled datasets is very challenging, particularly in the medical field, we employ self-supervised learning to exploit unlabeled data for learning effective visual representations for ophthalmic imaging data, in particular Fundus photo. We show the effectiveness of our approach by comparing our work against fully supervised approaches and pretrained network on non-medical data. We also assess the translation capacity of training with real-world data versus standardized data.

1.4 Classification for Small Datasets

Deep neural networks heavily rely on a large amount of data to show their full potential in modeling and solving problems. The generalization performance of deep learning models tends to degrade when trained with a small dataset. However, in real-world clinical settings, obtaining labels is very expensive, challenging, and requires the knowledge of domain experts. There are several fields in which acquiring a large amount of data is not even feasible, e.g. diagnosis of rare diseases. Hence, despite the exponential growth of deep learning algorithms in medical imaging applications, learning from extremely small datasets remains challenging. In Chapter 5, we aim to address the problem of learning from small datasets for the classification task. We present a novel framework, called CvS, that harnesses the power of segmentation to learn from a small dataset by incorporating a local dense loss and high-bias shape prior to the

learning process. As most classification datasets do not have segmentation labels, we employ two simple approaches, binarization and label propagation to obtain segmentation labels for the whole dataset. The label propagation method allows us to learn a preliminary model from a small subset of the dataset that is manually segmented and use this model to propagate the segmentation labels to the rest of the dataset.

CHAPTER 2

REAL-WORLD MULTI-MODAL LONGITUDINAL IMAGING DATASET FOR OPHTHALMIC APPLICATIONS

(This chapter was previously published as Mojab, N., Noroozi, V., Aleem, A., Nallabothula, M. P., Baker, J., Azar, D. T., Rosen- blatt, M., Chan, R. V. P., Yi, D., Yu, P. S., and Hallak, J. A.: **I-oda, real-world multi-modal longitudinal data for ophthalmic applications**. In Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Tech- nologies, BIOSTEC 2021, Volume 5: HEALTHINF, Online Streaming, February 11-13, 2021 , eds. C. Pesquita, A. L. N. Fred, and H. Gamboa, pages 566–574. SCITEPRESS, 2021 [1].)

2.1 Introduction

The past decade has witnessed an exponential growth in deep learning based applications in medical imaging [5–8]. The promising success of deep learning algorithms in computer vision, has motivated the immense growth of deep learning based modeling in medical imaging serving multiple purposes and addressing various problems ranging from classification to progression prediction or segmentation [9–16]. Most of the existing research studies focus on modeling and problem solving for medical imaging in artificial setting where the data is standardized and mostly collected from multi-center clinical trials. However, such data does not effectively represent the characteristics of real-world clinical data.

The lack of datasets that captures the true aspects of real-world data pose a major challenge for further improvements in medical computer vision and successful translation to real-world clinical settings. In this chapter, we mainly focus on addressing the above problem in the context of ophthalmic imaging domain. The current publicly available datasets that are commonly employed in the literature [18–21] suffer from five main limitations: (1) a small number of patients and imaging data, (2) data is standardized and mainly collected from artificial settings such as multi-center clinical trials, (3) lack of longitudinal data that contains imaging data at different time points for the same patient, (4) lack of multiple image modalities that allows studying the disease from different views of data , and (5) the potential risk of spectrum bias which emanates from the lack of diversity in population characteristics, such as sociodemographic and disease severity levels, that does not adequately represent the wide spectrum of patients in real-world clinical settings.

Different from the standardized dataset, real-world data is very complex, noisy, lacks standardization, and characterized by variability in quality, machine types, settings, and sources. Therefore, with standardized dataset not representing real-world data effectively, the translation capacity of models that are developed on the assumption of these standardized data, becomes limited. Hence, building a research-oriented medical imaging databank that reflects the key aspects of real-world data is imperative to advance the research in medical computer vision and improve generalizations and clinical translations.

With imaging technology advancing significantly over the past decades, the amount of generated imaging data is growing at significant rate. However, this data lacks standardization

and structure, collected from multiple heterogeneous settings, and lacks ground truth labels which pose major challenges on creating a large-scaled dataset from real-world data. Some of the major challenges include but not limited to: (1) limited access to original raw data due to Health Insurance Portability and Accountability Act (HIPAA), patient privacy, and ambiguity in data ownership, (2) data sources spread across multiple heterogeneous settings with very limited information on the data description, collection process and data integration across different sources, (3) lack of ground truth labels and standardization, (4) obtaining labels is very expensive, time consuming and requires domain knowledge of medical experts, and (5) complete anonymization for the whole data which is a very complex process due to the lack of consistent structure in the data.

Motivated by these challenges, we aim to create a medical imaging dataset for ophthalmic imaging applications from real-world data aiming to address the lack of real-world data and three core research problems in medical computer vision: (1) Advancing medical computer vision and machine learning-based applications in medical imaging and particularly in ophthalmology. (2) Providing an infrastructure to enhance generalizations and the translational capacity of deep learning based applications across different clinical settings. (3) Understanding disease progression trends across various ophthalmic diseases, and the variability among populations and the severity spectrum.

The Department of Ophthalmology and Visual Sciences at the Illinois Eye and Ear Infirmary of the University of Illinois at Chicago (UIC) is equipped with a rich collection of imaging data from a diverse patient population who received care over the past decade. This data contains

millions of raw unlabeled images and their metadata sitting across multiple sources with very limited information regarding the structure, label and data integration across different sources. We aim to develop an infrastructure that allows us to collect, preprocess, annotate, anonymize, and integrate the data from different data components across various settings at UIC to create a real-world ophthalmic imaging dataset.

In this chapter, we introduce a longitudinal multi-domain and multi-modal imaging dataset for ophthalmic imaging applications, called the Illinois Ophthalmic Database Applications (I-ODA). We present an efficient pipeline to collect, annotate, anonymize, and integrate the data. The dataset release is pending legal approval. To the best of our knowledge, this chapter is the first work attempting at creating a large-scale medical imaging dataset from a real-world data for ophthalmic imaging application.

Our dataset is characterized by five main key points: (1) more than 3.5 million image instances grouped into a diverse set of practical image modalities providing a comprehensive multi-view dataset for ophthalmic applications, (2) longitudinal imaging data for patients who received continuous care at one academic medical center over multiple time points, (3) a mixture of data from multiple imaging devices representing a multi-domain data, (4) a diverse patient population with various demographic background, and (5) a broad disease spectrum across multiple ophthalmic diseases. The unique properties of our dataset capture the characteristics of a real-world clinical data from different aspects each serving multiple purposes in ophthalmic imaging applications. I-ODA can provide an ideal benchmark for validation studies

and translations to patient care settings enabling breakthroughs in medical computer vision applications.

The rest of the chapter is organized as follows. We start by reviewing the data components that are used to create this dataset in section 2.2 and the challenges regarding data collection from each component. Then we propose our solution to address these challenges and how to integrate different data components into a structured imaging and relational database in Section 2.3. In Section 2.4 we discuss how to anonymize the dataset in regard to respecting the patient privacy. Section 2.5 provides a comprehensive information on characteristics of our dataset and discuss the importance of each of these properties from different aspects. Then we conclude the chapter in Section 2.6.

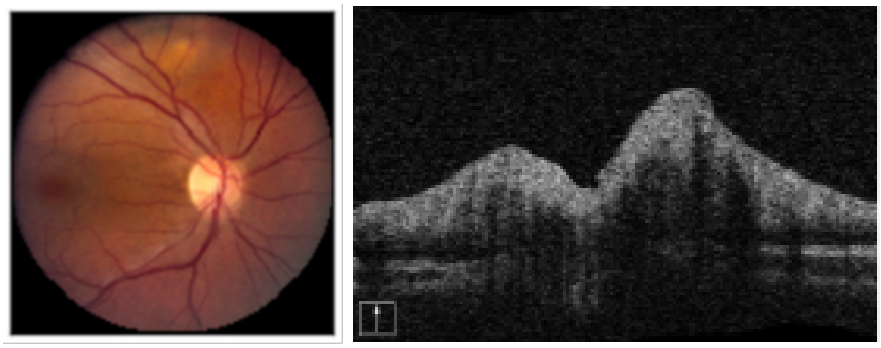
2.2 Database Atlas and Components

The Institutional Review Board (IRB) of the University of Illinois at Chicago approved the creation of the I-ODA databank. Each project that utilizes the I-ODA dataset will undergo additional review by the IRB to ensure patient privacy and protocol adherence. The research to build the I-ODA dataset was conducted in accordance with the requirements of the Health Insurance Portability and Accountability Act (HIPAA) and tenets of the Declaration of Helsinki.

The I-ODA dataset was created from imaging and clinical data belonging to the patients who visited the Illinois Eye and Ear Infirmary of the University of Illinois at Chicago (UIC) over the course of 12 years. The original data resides across three main sources: (1) Image files, (2) a SQL database, and (3) the University of Illinois Hospital and Health Sciences System (UIH) billing system.

2.2.1 Image Files

The image files are residing on an in-house server that maintains ~ 4.5 million raw images belonging to $\sim 45K$ patients. The raw image files are organized in a hierarchical structure sorted by Medical Record Numbers (MRNs), corresponding exam sessions, and image files residing on the in-house server that is connected to an image management system. The image files are generated by multiple imaging devices in the form of either a raw image of the eye or an analysis report. During each visit, patients can undergo multiple imaging test sessions for each eye. Based on the preliminary diagnosis identified by an ophthalmologist, photos representing different structures can be taken from multiple angles in each imaging test session. In the rest of this chapter, we refer to these images as "image modalities" which can be generated from different imaging devices. For example, a patient may require Fundus imaging, a photo of the posterior part of the eye, or Optical Coherence Tomography (OCT) imaging, which can represent high-resolution cross-sectional images of the retina. Fundus photos or OCT images are referred to as two types of imaging modalities. A sample of OCT and Fundus photos are illustrated in Figure 1. The modality of the image and the number of images taken per exam session could vary for each patient depending on the preliminary diagnosis. All the image files are originally stored in $.j2k (JPEG2000)$ file format with a broad range of image resolutions. All image files are unlabeled and are assigned with random file names that do not reveal any information regarding their modality, i.e. OCT or Fundus.



(a) Fundus (b) OCT

Figure 1: Illustration of sample images from Fundus modality and OCT modality.

The lack of consistent structure, different image resolution from devices with different setting and most importantly the absence of defined image modality and ground truth label for image files, pose major challenges to the first step of creation of labeled dataset.

2.2.2 Metadata

The metadata information that is used for the purpose of creation of our dataset is sitting across two main data components, a SQL Server database and the University of Illinois Hospital and Health Sciences System (UIH) billing system.

The SQL Server database consists of a collection of comprehensive information including but not limited to, patient demographics and their corresponding exam sessions, images taken in each session, and the imaging device generating the image files. The data reside across more than 50 tables in our SQL Server database. Although the SQL database contains a rich

collection of information, there is no descriptive information regarding the contents, structure or the purpose of each table's attributes and contents. There is also no constraint defined for the attributes constituting each table and hence no relation has been established across different tables.

These limitations can result in invalid, missing, and duplicate data records across the tables making the process of finding and extracting the data from the many SQL tables very challenging.

The UIH billing system contains a comprehensive information regarding hospital charges and diagnosis. For the purpose of this work, we extracted part of the data that encompasses the information regarding ophthalmic and non-ophthalmic diagnoses, interventions (clinical procedures), and patient demographics. This system is equipped with a billing report and a dashboard interface that allows us to retrieve hospital charges for limited number of data records at a time.

Due to manual entry from the imaging device interface and subjective errors both of the data components, SQL and UIH, are prone to noise and errors. Moreover, the lack of integration among the data components could result in inconsistent information across the data sources that needs to be addressed.

2.3 Methodology

The absence of defined modalities for the image files, lack of integration among the three sources of data and the noise emanating from human errors, pose major challenges to creation of labeled and structured dataset. To address these challenges, we create the I-ODA dataset

by designing a pipeline that is composed of three main phases each utilizing the data from one of the existing three data sources:

- Image modality tagging that assigns a modality tag to each image in the hierarchy of image files.
- Metadata annotation that extract and filter the valid data from the SQL database and connect it to the image files.
- Disease annotation that utilize the data from billing system to annotate image files and their metadata with the corresponding diagnosis.

Eventually all three sources of data are integrated creating the final dataset that is composed of two main data components, image files and relational database.

2.3.1 Image Modality Tagging

The modality tagging component of our pipeline consists of three steps, (1) drafting a set of potential image modalities, (2) yielding a set of prototype images per group of image modality, and (3) given the set of modalities and their prototype images, tagging each image with the proper modality.

2.3.1.1 Image Modality Selection

For the purpose of this paper, image modalities are defined as the most common imaging types used in ophthalmology. The selected set must encompass all representative modalities relevant to ophthalmic imaging applications and its diagnostic usage.

Each image is generated by one of the imaging devices that are used at the Illinois Eye and Ear Infirmary at UIC. To derive the preliminary set of modalities, we utilize the imaging device characteristics. Each imaging device is responsible for generating certain range of image modalities. However, this assumption might be violated in a few cases. Moreover, the set of image modalities generated by each imaging device is not necessarily exclusive. For example, two different modalities, Fundus and OCT, can be generated by three different devices. As the imaging devices do not necessarily generate one modality of images, they cannot be solely used for selecting the relevant modalities but can be further utilized as auxiliary information to narrow down the potential candidates.

One of the information that the SQL database contains is the list of all devices that are used to generate the image files that the UIC clinic has access to. Then, we utilized the device characteristic to draft a set of all potential image modalities generated by each device. Next, we selected random subset of images from each imaging device. Given the preliminary set of modalities, we manually reviewed images in each subset and selected the relevant modalities. We further reviewed the obtained modalities from each subset to potentially merge the relevant ones into one group. For instance, for images illustrating analysis reports containing OCT and Fundus images, one image modality referred to as "OCT Report" was chosen to represent both of these images. This step was repeated multiple times to achieve the final set of the most common and practical modalities which was further reviewed by ophthalmologists. This procedure helped us to keep the specificity level of each modality relevant to its diagnostic use in ophthalmology and enabled a practical collection of image modalities with a reasonable amount

of instances per modality. The final list contains 12 image modalities that are commonly used for ophthalmic diagnostic usage.

2.3.1.2 Image Prototype Selection

Images belonging to each modality can vary in terms of color, shape, and resolution but they are all to be considered as various members of the same modality. For instance, all varieties of Fundus images including square or circular shaped or black and white or colored should be tagged as one image modality named Fundus illustrated in Figure 2. Thus, selected prototype

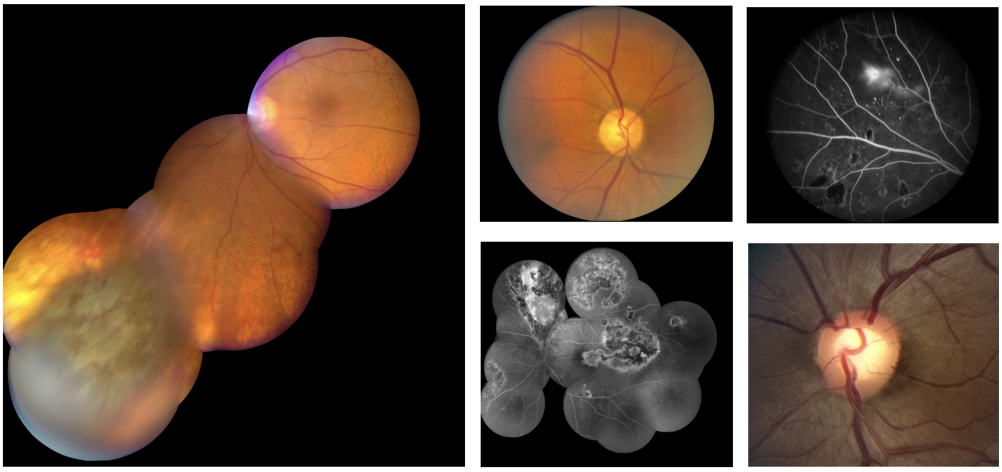


Figure 2: Illustration of modality spectrum for Fundus modality.

images for each group of modality must form a representative set of the whole spectrum of images belonging to that imaging modality. To select a set of representative image prototypes

per modality, we utilized the characteristic of imaging device and a similarity-based method. Given the set of modalities obtained from Section 2.3.1.1, we drafted a set of possible imaging devices that can generate each of the image modalities. Next, we selected a random subset of images from each device for each modality. This resulted in the preliminary set of prototypes for each modality group. To further refine the preliminary set of prototypes, we selected a random subset of images including all modalities. Then we employed a similarity-based method which will be elaborated on further in Section 2.3.1.3 to tag the selected subset of data by assigning the modality of their nearest neighbor from the prototype images in terms of euclidean distance. We then manually reviewed the results and analyzed the miss-classified images according to the characteristic of the members of each modality group. If the miss-classification occurred due to the absence of that particular image variation in its corresponding set of image prototypes, that image variation was added to its corresponding prototype set. We repeated this step multiple times each time augmenting the set of prototypes if necessary until we reached a negligible error for each modality. This step resulted in a final collection of 253 prototype images across 12 image modalities.

2.3.1.3 Tagging

Given the set of modalities and image prototypes, we propose a pipeline that takes the raw image with undefined modality as input and achieves the modality tag in two sequential steps.

- The first tag is achieved by employing a similarity-based classification method by comparing each image to the collection image prototypes.

- The obtained tag is verified by exploiting characteristic of imaging devices. The overall pipeline network is illustrated in Figure 3.

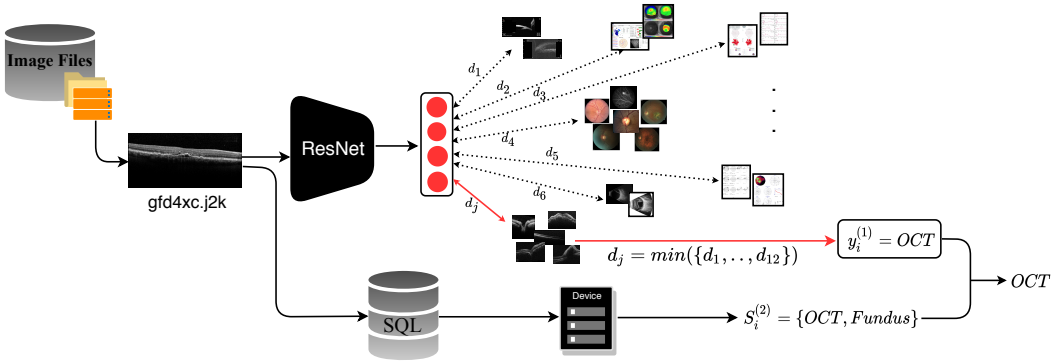


Figure 3: The illustration of overall pipeline for image modality tagging.

Similarity-based Classification: Suppose we have a dataset with N image instances and a set of M modalities. To tag each image with one of the M given modalities, we first employ a pretrained Convolutional Neural Network, ResNet-50, to extract the features for each image in the dataset and the set of prototype images. Suppose the dataset is denoted as $\mathcal{D} = \{x_1, ..., x_N\}$ where $x_i \in R^k$ represents the feature vector and k is its dimensionality. Given the M image modalities, we defined the set of prototype images as $\mathcal{V} = \{v^{(p)} | p = 1, .., M\}$ where $v^{(p)} = \{y_1^{(p)}, ..., y_{I_p}^{(p)}\}$, $y_{I_p}^{(p)} \in R^k$. $v^{(p)}$ represents the set of image prototypes for the modality group p and I_p denotes the number of instances in modality group p.

We aim to tag the images from the set $\mathcal{D}$ by assigning its nearest neighbor from the set $\mathcal{V}$ in terms of euclidean distance $j_p = argmin_{j_p} \|x_i - y_j^{(p)}\|$, $y_j^{(p)} \in \mathcal{V}$, $j = 1, ..., I_p, p = 1, ..., M$. To further ensure that the obtained minimum distance for the input image is reasonably chosen, we picked a threshold for each modality group by investigating the reasonable distance range among its image members. If the minimum distance achieved by a euclidean measure matched the threshold, we assigned the tag for the input image x_i by extracting the corresponding modality p associated with the index j_p in $\mathcal{V}$ denoted as $y_i^{(1)} = \mathcal{V}[I_{p-1} + j]$. We then applied the similarity-based method by comparing the images in $\mathcal{D}$ and the prototype images in $\mathcal{V}$ corresponding to all the 12 image modalities. The modality group of the nearest prototype image was chosen as the modality tag of the input image. The final set of modality tags achieved from this step is denoted as $\mathcal{Y}^{(1)} = \{y_i^{(1)} | i = 1, ..., N\}$, (depicted as $y_i^{(1)} = OCT$ in Figure 3).

Modality Verification: To validate the modality tag achieved from the first step, we narrowed down the possible set of modality tags for each image by utilizing its corresponding imaging device. We considered three subsets of data according to their corresponding imaging devices and the range of image modalities generated by each device, (i) images associated with devices that are responsible for generating only one type of imaging modality, (ii) images associated with devices that generate a specific range of imaging modalities (up to two modalities), and (iii) images associated with devices that their range of potential generated image modalities is not available in our data.

Given these three groups of subsets, we assigned each image in each subset to its possible set of modality tags according to its corresponding imaging device extracted from the SQL database. The first group of images which constituted $\sim 12\%$ of the data, were tagged with the one image modality generated by its corresponding imaging device. The second group which constituted $\sim 78\%$ of the data, was assigned with a set of potential modality tags according to their corresponding devices. The third group of images which constituted less than 1% of the data was assigned with an unknown tag. The set of modality tags obtained from each of these three groups of images is denoted as $\mathcal{Y}^{(2)} = \{S_i^{(2)} | i = 1, ..., N\}$ where $S_i^{(2)}$ represents the set of potential image modalities for the input image x_i (depicted as $S_i^{(2)} = \{OCT, Fundus\}$ in Figure 3).

Given the label sets $\mathcal{Y}^{(1)}$ and $\mathcal{Y}^{(2)}$, the final tag is assigned to each image if $y_i^{(1)} \in S_i^{(2)}$ for $i = 1, ..., N$ (depicted as OCT in Figure 3). Otherwise, it is assigned as unknown for further manual review and investigation.

2.3.2 Metadata Annotation

To extract the metadata corresponding to the image files we employed the data from the SQL database. First, we reviewed the contents of each table and its set of attributes to retrieve those that are relevant to creation of our dataset. Overall we isolated four tables that maintain information regarding patient, image files, exam session, and imaging device. In each table we only keep the relevant attributes.

Next, we apply a series of pre-processing steps to filter out the invalid data records such as invalid MRNs, duplicate, and missing records across the related tables. The main preprocessing

steps include filtering out invalid MRNs, duplicate MRNs with different data records, and missing data records across the relevant tables. Further, we exclude inconsistent data records across the two data sources, image file hierarchy and SQL database using the list of directories an sub-directories in the image files hierarchy and their equivalence in SQL data, patient MRN and session Id. The patient and image file tables originally contains $44,460$ patients and $4,477,634$ image files, respectively. Applying the preproccessing steps results in exclusion of $\sim 8\%$ of the data. Eventually, we joined all the tables into one metadata file using the mutual attributes among the selected four tables.

2.3.3 Diagnosis Annotation

Given the validated metadata file and patient MRN from Section 2.3.2, we utilize the billing reports from UIH system to retrieve ophthalmic and non-ophthalmic diagnoses and interventions for each patient. The interventions are referred to as any surgical or invasive outpatient or hospital procedure. Next, similar to the Section 2.3.2, we apply a series of preprocessing steps that resulted in removing $\sim 12\%$ of the data. Then we exclude the inconsistent data records across billing data and metadata file which resulted in exclusion of another $\sim 6\%$ of the data. Finally, we merged the billing data with the metadata file obtained from the SQL database. The final file contains $33,876$ patient and $3,668,649$ image files which are annotated with their corresponding metadata, diagnoses, and interventions.

2.3.4 Data Integration

At last, we constructed a relational database integrating the data from all the data sources, image files, metadata, and diagnoses. This database constitutes of 6 tables including patient

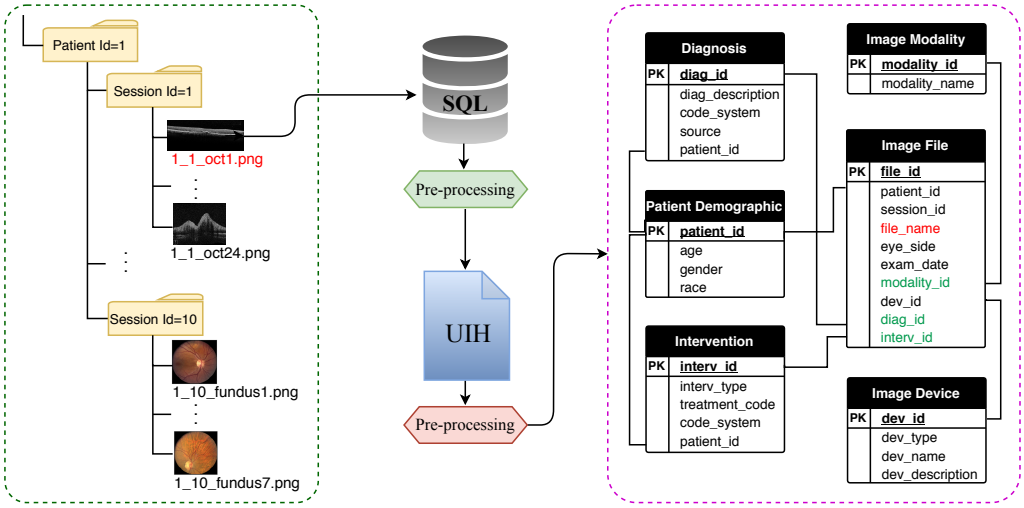


Figure 4: Illustration of the pipeline network for disease annotation and integration of the three data sources, image file hierarchy (depicted in green bounding box), SQL data, and billing data.

demographic, image file, diagnosis, intervention, imaging device, and image modality. The tables are connected through a primary key (PK) and a foreign key constraint defined for each table. Each table consists of a set of attributes with its metadata information serving its purpose in our dataset. The data integration and relational database schema (depicted in a pink bounding box) is illustrated in Figure 4. As Figure 4 shows, each image (depicted in red) is annotated with its corresponding image modality, diagnosis, and interventions (depicted in green).

2.4 Data Anonymization

Data anonymization is the process where patient identifiers are irreversibly removed for patient privacy protection, prohibiting any direct or indirect identification. According to the HIPAA regulations, sensitive patient information should be protected by being properly anonymized before being used for any research purposes. Data anonymization in the context of our work would result in a complete anonymized dataset across both data components, image files, and the relational database.

2.4.1 Image Anonymization

The image members in each of the 12 modality group in our dataset can vary in terms of style, resolution, and location of identifiable information that needs to be masked out. The extensive range of variability among images poses a major challenge on anonymization for such a large amount of data. To address this challenge, we employ a K-means clustering method to derive a set of categories for each of the 12 modalities where the images in each category are the most similar ones in terms of style, resolution, and location of identifiable information. To choose the initial number of clusters for each modality group, we first randomly selected a subset of 200 images from each modality and manually reviewed and analyzed the selected subsets. We further applied a set of various imaging filters, including spatial/geometric, resolution, appearance, and color, to achieve a more fine-grained categorization for each of the categories obtained from the initial clustering. The set of filters were chosen to be relevant to the type of images belonging to each modality.

Next, we divided the obtained categories into two groups based on the consistency level of the location of identifiable information. For the first group of categories that maintain a consistent pattern in terms of location of identifiable information across their image members, we generated a location-based masking filter. The masking filter are specific to each category and are further employed to mask out the part of the image that contains the identifiable information. The second group of categories for which the location of identifiable information varied across their image members, we combed through the data and manually removed the sensitive information. Eventually, the data was reviewed by two persons to anonymize any missed data to ensure complete anonymization of our image data. The fine-grained categories were created merely for the purpose of anonymization. After accomplishing the anonymization process for all the image files, the categories were discarded and only the 12 main modalities that were obtained in Section 2.3.1.1 were kept.

2.4.2 Metadata Anonymization

To de-identify the metadata, first, we extract the set of sensitive attributes including the patient MRN, first and last name, date of birth, and exam session date. The patient first and last names were removed, and the MRNs were replaced by randomly generated numbers. To keep the longitudinal nature of the date of birth and exam session dates attributes, the date of births were replaced by patient's age and the date of exam sessions were replaced by subtracting the date of births from the date of exam sessions. To integrate the anonymized metadata with the anonymized image files, the patient and exam session directories in the hierarchical structure of image files were renamed to the anonymized patient ids and exam session ids respectively.

2.5 Dataset Characteristic

I-ODA dataset captures different characteristics of a real-world clinical data from different perspectives. These unique properties distinguishes our dataset from existing ophthalmic imaging dataset. In this section we explore these characteristics and show how each can potentially address some of the major challenges in medical imaging. We further showcase their contributions in improving clinical translation enabling versatile computer vision applications in medical imaging and ophthalmology in particular.

2.5.1 Dataset Statistic

As of now, the I-ODA dataset contains $3,668,649$ images and $230,923$ exam sessions across 12 image modalities of $33,876$ individuals from the Department of Ophthalmology and Visual Sciences at the Illinois Eye and Ear Infirmary of the University of Illinois at Chicago for eye care. The set of image modalities includes {Optical Coherence Tomography (OCT), OCT Report, Fundus, Humphrey visual field (HVF), Ultra-sound, Ultrasound Report, B-Scans, Corneal Topography, External image (slit lamp), Intraocular Lens master calculation report (IOL), Optical Response Analyzer report, ERG report}.

Among the 12 image modalities, 6 modalities, Fundus, OCT Report, OCT, HVF, B-Scan, Corneal Topography, constitute 98% of the imaging exam sessions. A snapshot of samples from these 6 modalities is illustrated in Figure 5. As Figure 5 suggests, each image modality encompasses a spectrum of different varieties of its image members.

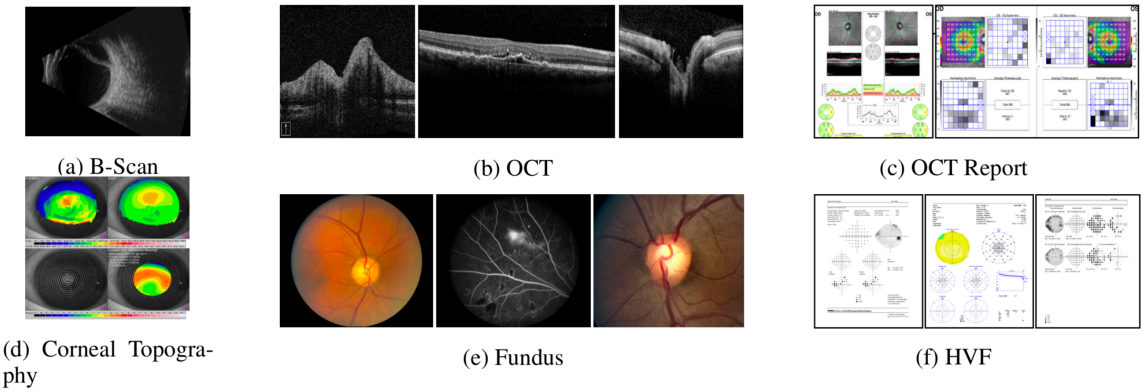


Figure 5: A snapshot of samples from 6 major image modalities in I-ODA dataset.

2.5.2 Dataset Components

The I-ODA dataset is composed of two main data components integrated effectively to represent a structured ophthalmic imaging dataset, as shown in Figure 4:

(1) Anonymized image files that are tagged with their corresponding modality. Images are converted to *.png* format and stored in a hierarchical structure where the highest level represents a patient directory followed by its corresponding exam session and finally the imaging files that reside on the lowest level of the hierarchy. The patient and exam session directories correspond to the anonymized patient ids, and session ids from the metadata. The image file names are formatted as "patientId_sessionId_modality" format.

(2) A relational database that constitutes of 6 tables representing patient demographics, image files, diagnoses, interventions, imaging devices, and image modalities integrated through primary and foreign key constraints. As Figure 4 suggests, the corresponding patient metadata,

diagnosis, and intervention (depicted in green in the "Image File" table) for each image file (depicted in red in image file hierarchy on the left) can be easily retrieved from the tables in our relational database.

2.5.3 Modality and Domain

The I-ODA dataset comprises 12 different image modalities. The imaging data collected in the process of ophthalmic disease diagnosis can include multiple imaging sessions with different image modalities per patient visit. This would result in a rich collection of longitudinal imaging sessions across different image modalities for ophthalmic applications. In real-world clinical settings, for a physician to make an accurate diagnosis, they often need to look at multiple views of the data such as different imaging tests. A summary of the I-ODA dataset showing the 6 major image modalities and the number of exam sessions per modality is illustrated in Figure 6.

The availability of a vast number of imaging sessions across a comprehensive set of multiple imaging modalities, allows us to study the disease pattern from multiple sources of data which could potentially lead to more accurate diagnosis and treatment.

Moreover, the image data in I-ODA is generated from more than 30 imaging devices forming a multi-domain dataset, where domain in here is defined as the imaging device. This important property reflects the true nature of a real-world dataset, which can include a mixture of data distributions collected from different domains. Given that clinical care often involves complex multi-domain data, I-ODA could provide a potential benchmark for validation stud-

ies towards improvement in generalizations and translations of machine learning based models across different clinical settings.

2.5.4 Patient Population

I-ODA contains a rich collection of imaging data and metadata from a diverse set of patients with various demographic backgrounds including, ethnicity, race, age distribution, and location. A cross-tabulation analysis of patient gender and age from I-ODA is depicted in Figure 7.

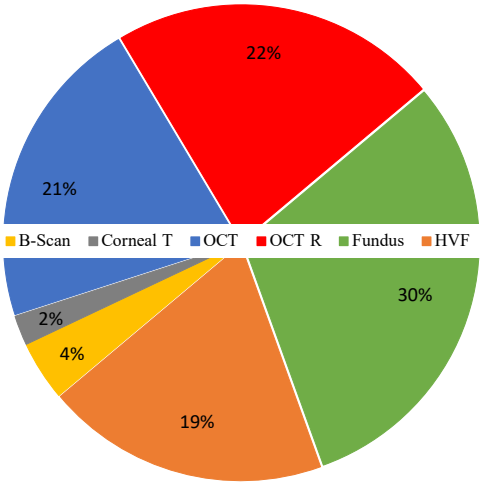


Figure 6: A snapshot of number of imaging sessions for modalities Fundus, OCT Report(OCT R), OCT, HVF, B-Scan, Corneal Topography(Corneal T)

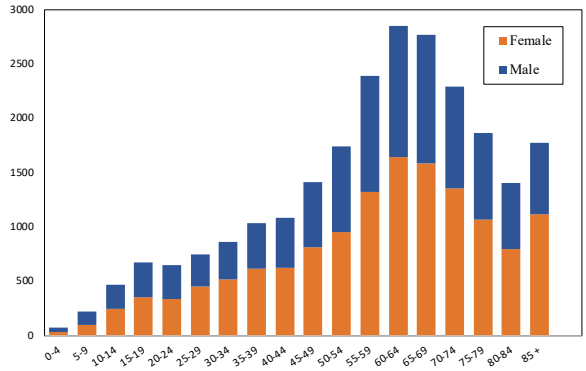


Figure 7: Illustration of a cross-tabulation analysis of patient gender/age population distribution in I-ODA dataset.

As Figure 7 suggests, the patient population in I-ODA has a comparable distribution between females and males across a wide age range. This will allow the development and validation of machine learning based algorithms that are a true representation of the patient population and hence improve the generalization capacity in modeling and problem solving.

2.5.5 Longitudinal Disease Spectrum

I-ODA contains an extensive set of patient visits from one academic institution having received imaging tests across multiple time points. Additionally, I-ODA encompasses a comprehensive collection of imaging data for various ophthalmic diseases including but not limited to diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma. Having the longitudinal nature of imaging sessions from patient visits over time allows us to capture a broad disease spectrum. Figure 8 represents the severity spectrum of Fundus photos taken at different stages of one of the ophthalmic diseases, diabetic retinopathy (DR). NPDR represents Non-proliferative Diabetic Retinopathy, ME represents Macular Edema, and PDR represents Proliferative Diabetic Retinopathy. This property allows us to study the progression

Figure 8: Illustration of the disease spectrum of DR with Fundus photos taken at different stages.

trends across different ophthalmic diseases. Having a rich collection of different ophthalmic diseases enables studying diseases both in isolation and correlation with each other. Additionally, the availability of non-ophthalmic diseases in our dataset allows us to identify common biological and epidemiological mechanisms by studying the correlation among ophthalmic and non-ophthalmic diseases [22]. Additionally, the longitudinal diseases spectrum and diverse patient population mitigates the risk of spectrum bias that emanates from the lack of diversity in population characteristics such as sociodemographicand disease severity level.

2.6 Conclusion

In this chapter, we study the problem of creating a medical imaging dataset from real-world data aiming to address the lack of real-world clinical data and improving upon clinical translation. Creating a labeled and structured dataset from real-world data that effectively captures the characteristics of real-world data is significantly challenging. The reason is because the data is unlabeled, lacks standardization and structure, and collected from multiple heterogeneous settings with no information regarding data components, data integration, labels, and anonymization process.

To address these challenges, we propose an infrastructure to collect, process, label, anonymize and integrate the data from different sources into a unified imaging dataset, called I-ODA. I-ODA is a multi-domain longitudinal data containing a diverse patient population providing an ideal benchmark for validation studies and clinical translation. We present the unique characteristics of I-ODA that distinguishes our dataset from other ophthalmic imaging dataset. These characteristics include but not limited to, a large amount of data across a diverse collection

of image modalities and domains, diverse patient population who received continuous care at the Department of Ophthalmology and Visual Sciences at the Illinois Eye and Ear Infirmary at UIC, and longitudinal imaging data.

We showcased that I-ODA dataset with its unique characteristics serves a wide range of purposes in medical imaging applications. The broad disease severity spectrum, demographics, and interventions represent a diverse patient population with different outcomes. Applications on imaging data may particularly help with delineating structural changes that indicate future vision loss. This will improve our understanding with regards to the different progression patterns among different patient populations. Moreover, the diverse collection of longitudinal ophthalmic and non-ophthalmic diseases will allow us to study potential correlation patterns among different diseases. The mixture of multiple domains, emanating from different imaging devices alongside the multiple image modalities, and longitudinal disease spectrum, reflects characteristics of real-world clinical data which is of great importance for clinical translations. I-ODA dataset can provide a potential benchmark for validating deep learning models and their clinical applicability to targeted populations.

CHAPTER 3

MULTI-TASK LEARNING FOR CLINICAL INTERPRETABLITY

(This chapter was previously published as Mojab, N., Noroozi, V., Yu, P., and Hallak, J.: **Deep multi-task learning for interpretable glaucoma detection**. In 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI) , pages 167–174. IEEE, 2019 [2].)

3.1 Introduction

Success of deep learning algorithms in computer vision applications, has motivated the immense growth of deep learning based application in medical imaging. These advances has inspired many research studies aiming to address various problems from diseases classification to localization of imaging biomarkers. The advances in automated decision making in medical field can help with large-scale screening which could potentially lead to early diagnosis and prevent the diseases from further progression. Motivated by these advances, ophthalmic imaging applications, in particular detection of blinding diseases, have received considerable attention in the last decade. Early detection of blinding diseases can serve a great benefit to preserving the vision. With deep learning based applications growing exponentially in medical imaging, there has been a growing concerns regarding the clinical interpretability of the model's solution to the task. Clinical interpretability plays an important role accurate and reliable decision making in clinical settings. However, current research studies mostly disregard the transparency of the

model's solution to the task and merely focus on predicting the final output, e.g. predicting whether an input image is diseased or healthy. The lack of clinical interpretability limits clinical applicability of such methods. In this chapter, we aim to address the clinical interpretability in the context of real-world ophthalmic application, glaucoma detection.

In real-world clinical applications, indication of glaucoma is assessed by providing imaging biomarkers, optic disc and optic cup and clinical measurement CDR, in a Fundus photo. The general problem of glaucoma detection in ophthalmology is well studied in the literature. This problem is commonly formulated into a binary classification task which identifies the presence or absence of glaucoma directly from a Fundus image using a deep learning model. However, these methods disregard the clinical interpretability of the results which limits their clinical application in real-world setting.

There are some studies attempted at achieving clinical interpretability by locating imaging biomarkers, optic disc or optic cup, via segmentation methods. However, segmentation labels are very expensive and time-consuming to obtain which limits the application of such methods. Moreover, different from other fields of computer vision, locating biomarkers in a medical image requires the domain knowledge of medical experts.

On the other hand, classification labels are more accessible and easier to obtain. Therefore, it is desired that the larger amount of data with classification label can be efficiently utilized to mitigate the shortage of segmentation labels while benefiting from the interpretability of the results from segmentation methods.

In this chapter, we introduce a novel framework, called InterGD, to the above problem. We formulate the problem into a multi-task learning setting to exploit the complementary information across two modules, classification and segmentation. Different from existing methods, our approach, utilizes both data with classification labels and the data with segmentation label. The network is designed in a way that the classification module learns to predict glaucoma from the predicting disc and cup maps in the segmentation component. Therefore the segmentation component would benefit from the classification component that has access to enough labeled data, to alleviate the shortage of segmented data while achieving interpretable results. We show the effectiveness of our approach through comprehensive experiment on several datasets. To the best of our knowledge, this chapter is the first to address the problem of glaucoma detection in multi-task learning setting that achieves clinical interpretability.

The rest of the chapter is organized as follows. We start by a brief review on glaucoma, the imaging that are commonly used for glaucoma diagnosis and the imaging biomarkers that are indication of the disease in Section 3.2. In Section 3.3 we review related works on glaucoma detection. Then we introduce our method and problem formulation in Section 3.4 and model architecture in Section 3.4.2. Section 3.5 reports the experimental results and model analysis. Then we conclude the chapter in Section 3.6.

3.2 Preliminary Concepts

Glaucoma is a complex disease that gradually leads to optic nerve damage, resulting in progressive irreversible vision loss. Over 60 million people are diagnosed with glaucoma, encompassing more than 8 million cases with irreversible blindness [23]. The global incidence of

glaucoma is anticipated to increase up to 111.8 million by 2040 [24]. Vision loss in glaucoma happens gradually, where patients become symptomatic in advanced stages makes its diagnostic more challenging.

Ocular imaging is one of the main modalities used for glaucoma screening. Among the imaging modalities most commonly used for glaucoma, digital Fundus photos are heavily utilized for their noninvasive optic nerve head evaluation, allowing the visualization of the disc and cup areas. The assessment of the optic nerve head is based on measuring the optic disc and optic cup regions in Fundus photos and calculating the cup-to-disc ratio (CDR). Figure 9 illustrates a sample of glaucomatous and non-glaucomatous optic nerve, with localization of the optic disc and optic cup regions. In Figure 9, the top row represents a glaucomatous Fundus of the eye and the bottom row shows a non-glaucomatous Fundus. The area of the optic nerve head is zoomed in for better visualization of optic disc and cup. CDR is calculated by dividing the cup height (in green arrows) over disc height (in bright blue arrows). As Figure 9 suggests, the CDR value increases, the risk of a Fundus become glaucomatous gets higher.

The assessment of CDR heavily relies on the expertise of ophthalmologists which not only limits the large-scale screening but also leads to less accessibility in remote areas, high variability among graders, and increase in the probability of human errors. Therefore, developing a system than can automatically identify the presence of glaucoma is essential and would greatly benefits the ophthalmology field in improving patient's health and preserving vision.

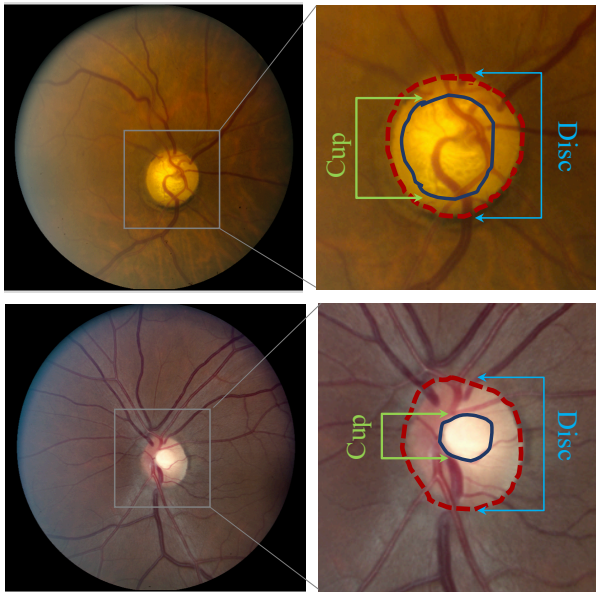


Figure 9: Samples of glaucomatous (top row) and non-glaucomatous (bottom row) Fundus photo from I-ODA dataset.

3.3 Related Works

The current approaches on glaucoma classification from a Fundus photo falls into two main categories: (i) Segmentation based approaches which utilize the CDR measurement to detect glaucoma. These approaches identify the presence of glaucoma by extracting the two imaging biomarkers, disc and cup regions, and measuring the CDR in Fundus photos [25–27]. (ii) Classification based approaches which predict the presence of glaucoma directly from a Fundud photo [28–30]. These approaches mainly employ Convolutional Neural Networks (CNNs) specifically designed for the dataset in use.

Segmentation based approaches generate interpretable results by localizing the optic disc and cup, and calculating the CDR to make the diagnosis. However, in real-world applications, the segmentation labels are very expensive and challenging to obtain. Further, the accuracy may get affected by the noise inherent in the manual search for Region of Interest (ROI) and thus leading to potential performance degradation. Additionally, focusing merely on the optic disc and cup regions, ignores the global visual context information in Fundus photos that may contain relevant information regarding the indication of disease. These challenges could potentially limits the applications of such methods in practice. On the other hand, acquiring labeled data for classification based approaches are much easier and more readily accessible. Different from segmentation-based methods, these approaches usually utilizes a global visual context by extracting features directly from a full Fundus photos to indicate the presence of glaucoma. However, these approaches suffer from the lack of interpretability that is imperative for improving clinical applicability.

Our approach leverage both data with classification labels and segmentation labels by formulating the problem in multi-task learning framework. The segmentation module in our model targets the clinical interpretability of the model. Different from standard multi-task learning, our model predicts glaucoma from the CDR assessment of predicted disc and cup from segmentation task. This design forces the model to learn how to predict glaucoma from the segmentation maps, and therefore it makes the segmentation parts to benefit from the part of the data with just classification label and thus address the shortage of labeled data in segmentation task.

3.4 Methodology

The propose model is composed of two main complementary components, segmentation and classification, integrated into a multi-task framework. The segmentation module focuses on localizing the regions of disc and cup in a Fundus photo. This component targets the transparency of the model's solution to the task as it provides the CDR measurement from extraction of imaging biomarkers, disc and cup area. The classification component focuses on detecting the presence of glaucoma from a full Fundus photo. This component aims to alleviate the shortage of labeled data in segmentation component. The network is designed in a way that the classification module learns to predict glaucoma from the obtained disc and cup maps in the segmentation component. Therefore the segmentation component would benefit from the classification component that has access to enough labeled data, to alleviate the shortage of labeled data while achieving interpretable results.

3.4.1 Problem Formulation

Before presenting our proposed approach, we first introduce the notations that will be used throughout this chapter. Suppose the training set is composed of two sets of data samples $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_s$. Let $\mathcal{X}_s = \{(x_k, d_k, c_k)|1 \leqslant k \leqslant N_s, d_k(i,j) \in \{0,1\}, c_k(i,j) \in \{0,1\}\}$ denote the part of our training samples having only segmentation labels for disc (d_k) and cup (c_k). This set does not contain classification label and it is used for training the segmentation module.

Let $\mathcal{X}_p = \{(x_k, y_k)|1 \leqslant k \leqslant N_p, y_k \in \{0,1\}\}$ denote the part of the training samples having only the classification label where y_i represent the binary class label, glaucomatous or

non-glaucomatous. This set does not contain segmentation labels and it is used to train the classification module. Our goal is to learn the following functions

$$g_d(z_k;\theta_d) = d_k \tag{3.1}$$

$$g_c(z_k;\theta_c) = c_k \tag{3.2}$$

$$h(d_k, c_k) = y_k \tag{3.3}$$

The two funtions $g_d(.)$ parameterized by θ_d and $g_c(.)$ parameterized by θ_c predicts the segmentation map of disc and cup respectively from the feature map z_k. Given the segmentation maps d_k and c_k, the function $h(.)$ predicts the probability of an input Fundus being glaucomatous. The function $f(x_k;\theta_f) = z_k$ parameterized by θ_f maps the input x_k to feature map z_k. All of the functions are modelled by deep neural networks. The schematic representation of InterGD is illustrated in Figure 10.

3.4.2 Model Architecture

The propose model is composed of two main complementary components, disc-cup segmentation and classification, formulated into a multi-task learning setting. The overall network architecture of the proposed model is illustrated in Figure 11. The network takes a full Fundus

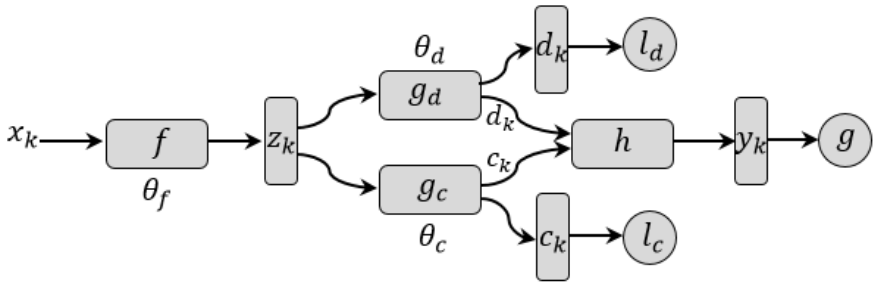


Figure 10: The schematic representation of InterGD. All functions f, g_c, and g_d are modelled by neural networks.

image as input and infers three main predictions: 1) segmentation map of the optic disc region, 2) segmentation map of the optic cup region and 3) class label (glaucoma or non-glaucoma).

3.4.2.1 Backbone Network

The main block of our framework that corresponds to function f obtains the representation z. This part of the model is built upon a U-Net architecture [31] that is composed of encoder and decoder paths. The encoder path consists of five stacked convolutional blocks each consists of two convolution layers with ReLU and linear activation followed by a batch normalization and nonlinearity of ReLU. The first four blocks are followed by a max pooling to downsample the resolution of a given input image. The decoder path is composed of four convolutional blocks each consist of up-sampling layers followed by batch normalization and nonlinearity ReLU. The up-sampling layers in the decoder path is concatenated to its corresponding encoder features

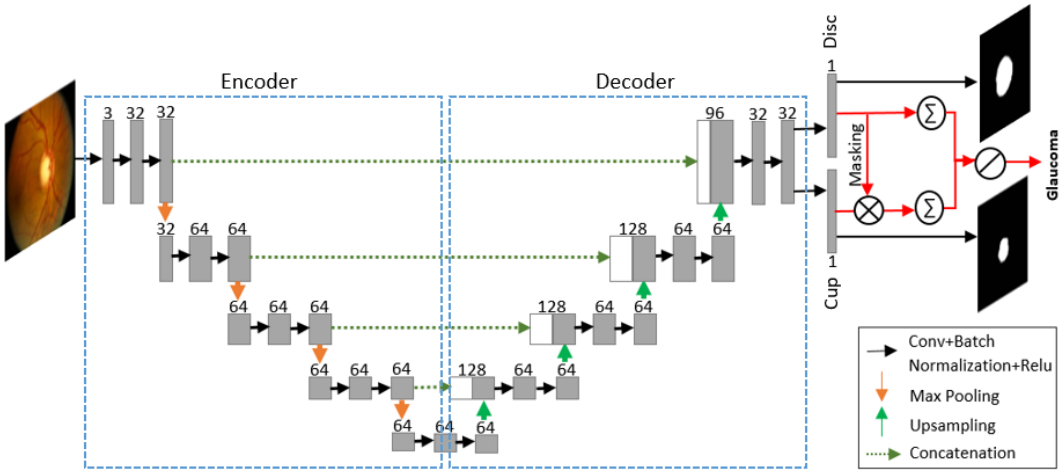


Figure 11: Illustration of overall architecture of InterGD for interpretable glaucoma screening.

maps. The main block of the model is shared among the three branches, allowing the model to leverage from the learned features in each module.

3.4.2.2 Disc-Cup Segmentation

The obtained feature map z is given to two separate heads. Then z is convolved with a 1×1 convolution and a sigmoid activation is applied to predict the segmentation maps of disc (function g_d) and cup (function g_c).

Given $\mathcal{X}_s$ and its corresponding label set $R \in \{D, C\}$ where D and C represent the cup and disc segmentation maps, the loss function for the two segmentation heads over all N_s samples is defined as follows:

$$\mathcal{L}_s(\mathcal{X}_s, \mathcal{R}) = \sum_{k=1}^{N_s} \ell_s(\hat{r}_k, r_k) \tag{3.4}$$

where $r_i \in \{d_i, c_i\}$ and $\hat{r}_k$ and r_k represent predicted and ground truth labels respectively. We employ dice-coefficient loss for the purpose of this task. Dice Similarity Coefficient (DSC) in our problem measures the spatial overlap between the predicted segmentation map and the ground truth label. The loss function $\ell_s(\hat{r}_k, r_k)$ for one sample is defined as follows:

$$\ell_s(\hat{r}_k, r_k) = -\log \frac{2\sum_{i,j=1}^{H,W} \hat{r}_{k,(i,j)} + \epsilon}{\sum_{i,j=1}^{H,W} {r^2}_{k,(i,j)} + \sum_{i,j=1}^{H,W} \hat{r}^2_{k,(i,j)} + \epsilon} \tag{3.5}$$

where H and W are the height and width of the input image, $r_{k,(i,j)}$ and $\hat{r}_{k,(i,j)}$ represent the ground truth label and predicted probability for ij-th pixel and $\epsilon > 0$ is a small number added to prevent division by zero. The overall loss used to train the segmentation module can then be defined as:

$$\mathcal{L}(\mathcal{X}_s, \mathcal{D}, \mathcal{C}) = \mathcal{L}_s(\mathcal{X}_s, \mathcal{D}) + \mathcal{L}_s(\mathcal{X}_s, \mathcal{C}) \tag{3.6}$$

where $\mathcal{L}_s(\mathcal{X}_s, \mathcal{D})$ represents the loss function for disc segmentation and $\mathcal{L}_s(\mathcal{X}_s, \mathcal{C})$ represents the loss function for cup segmentation.

3.4.2.3 Classification

The classification component identifies the presence of glaucoma in a given Fundus by utilizing the disc and cup regions obtained from the segmentation part. The resulting segmentation maps from the disc and cup branches are fused together and the assessment of CDR is used to predict the presence of glaucoma. This architecture forces the model to learn how to predict glaucoma from the segmentation maps, and therefore it makes the segmentation parts to benefit from the part of the data with just glaucoma label and thus address the shortage of labeled

data in segmentation task. Moreover, to ensure a correct alignment of cup region inside the disc area, we employed a masking technique in which disc map is applied on the cup map as a mask thorough an element-wise dot product. Then the area of the disc and masked cup are calculated by summation pooling over all the regions. The ratio of cup over disc area (CDR) can then be an indicator for detecting glaucoma. This part of the model estimates the function h.

Given $\mathcal{X}_p$, the loss function for the classification part is defined over the N_p samples as:

$$\mathcal{G}(\mathcal{X}_p, \mathcal{Y}) = \frac{1}{N_p} \sum_{k=1}^{N_p} g(\hat{y}_k, y_k) \tag{3.7}$$

where y_k and $\hat{y}_k$ are the ground truth and predicted labels respectively. The loss function $g(\hat{y}_k, y_k)$ for one sample is defined as the cross-entropy between the ground truth and the model's estimation as:

$$g(\hat{y}_k, y_k) = -y_k \log \hat{y}_k + (1 - y_k) \log(1 - \hat{y}_k) \tag{3.8}$$

3.5 Experiments

In this section, we conduct extensive experiments to examine the effectiveness and efficiency of InterGD on several datasets for glaucoma detection and clinical interpretability.

3.5.1 Datasets

We employed two datasets, DRISHTI-GS [20] and RIGA [19] for segmentation component and I-ODA [1] for classification task.

I-ODA The Illinois Ophthalmic Database Atlas (I-ODA) [1] has been created from imaging data belonging to patients who visited the Illinois Eye and Ear Infirmary of the University of Illinois Chicago (UIC). We isolated two subsets of data I-ODA-A and I-ODA-B with different number of images to assess the effect of data size on the segmentation module. These datasets do not have any segmentation labels. I-ODA-A contains 4997 glaucomatous and 5000 non-glaucomatous Fundus photos and I-ODA-B contains 1232 glaucomatous and 1267 non-glaucomatous Fundus. The non-glaucomatous Fundus are defined as being diagnosed with neither glaucoma nor glaucoma suspect with no potential damage to the optic nerve head.

DRISHTI-GS dataset [20] contains 50 full Fundus images with their corresponding optic disc and optic cup segmentation maps manually segmented by multiple human experts.

RIGA data set [19] contains 750 Fundus images and 4500 images with disc and cup contour manually marked on the image. The data is spread across three main datasets: Bin Rushed, Magrabia and MESSIDOR. We only employed two datasets Bin Rushed and Magrabia in our experiments. Each Fundus is marked by a couple of physicians. we count each segmented Fundus as an independent sample. Since RIGA dataset does not contain segmentation maps of disc and cup, we employed a two stage approach to extract the segmentation maps from the contours manually marked on the images. We first extracted the contours through morphology techniques and measuring region properties. Then we employed convex coordinates to extract the regions of disc and cup and their corresponding segmentation maps. We have isolated 540 disc and 540 cup segmentation maps belonging to 115 fundus images. We considered multiple segmentation maps corresponding to each image as different samples. Therefore, in total we

isolated 540 colored fundus images with their corresponding disc and cup segmentation maps. This dataset does not contain any classification labels.

Datasets DRISHTI-GS and RIGA are utilized for training the segmentation part of the proposed model while I-ODA is used for the classification task.

3.5.2 Baseline Methods

In order to demonstrate the effectiveness of our approach, we compare InterGD with three baseline methods in the context of disc-cup segmentation and two methods in the context of classification for glaucoma detection. We also adopt two versions of our models that are trained for a single task of either segmentation or classification.

U-Net Sevastopolsky et [26]: This work employs a modified U-Net consisting of encoder and decoder paths with skip-connections between corresponding layers in encoder and decoder paths. This approach is designed in a way that the disc and cup segmentation are generated sequentially. The network is first trained to obtain the disc segmentation map and then the image is cropped to the region of optic disc that is further being fed to the network to output the cup segmentation map.

CNN Zilly et [32]: This works employs an ensemble learning to learn convolutional filters. Entropy sampling is further utilized to select informative points and reduce the computational complexity.

InterGD-Seg: This model is an adopted version of InterGD where the classification module is disabled and the model is trained solely for a single-task of disc-cup segmentation with two heads for disc and cup.

CNN: The network architecture is mainly borrowed from [33]. The model consists of five convolutional blocks where each consists of two convolutional layers with ReLU and linear activation followed by batch normalization and max pooling layer. Finally a fully connected layer and sigmoid were applied to predict the class label.

InterGD-Pred: This model is another adopted version of InterGD in which the segmentation module is disabled to solely assess the performance of the classification task.

3.5.3 Training and Optimization

The training of InterGD consists of two alternating phases. In the first phase, the model is trained for n_s epochs using backpropagation with respect to objective functions Equation 3.1, and Equation 3.2. In the second phase, we train the model for n_p epochs using backpropagation with respect to the objective function Equation 3.3. Given $\mathcal{X}_s$ and $\mathcal{X}_p$, we optimize the model through Adam optimizer [34] for both phases and for a total of n epochs. For each epoch, samples are shuffled and training is done with mini-batches. Batch normalization [35] is employed after each convolution layer to improve the optimization.

3.5.4 Experimental Setting

A split of 80% and 20% is used for training and testing, respectively for all three datasets. Hyperparameters are selected through validation on a 20% randomly selected subset of the training data. After finding the appropriate values for parameters through validation, all the training dataset is used for the training.

The images from all three datasets were resized to $256 \times 256 \times 3$. Then, channel-wise normalization is applied on all the images. The batch size were selected from $\{16, 32, 64, 128\}$.

The number of training epochs was chosen from $\{1, 5, 10, 25, 50, 100, 200, 300, 500, 1000\}$ for segmentation and $\{1, 10, 20\}$ for classification. Eventually, the model was trained for $n_s = 50$ epochs and $n_p = 1$ when using I-ODA-A and DRISHTTI-GS as training datasets. The training was done for a total number of $n = 500$ times. The model was trained for $n_s = 20$, $n_p = 1$ and a total number of $n = 500$ when using I-ODA-B and DRISHTTI-GS as training datasets. When using the RIGA dataset for training, we set the values of $n_s = 5$, $n_p = 1$ and trained for a total number of $n = 300$ with I-ODA-A. We used the same hyperparameter setting for I-ODA-B except for the total number of training iterations being set to $n = 500$. The baseline model CNN is trained using the I-ODA-B dataset, for a total number of 500 training iteration. We used Adam optimizer with learning rates chosen from $\{0.01, 0.001, 0.0001\}$.

3.5.5 Performance Metric

The proposed model and baselines are evaluated in terms of precision, recall and F1-score for classification as follows.

$$Precision = \frac{TP}{TP + FP} \tag{3.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.10}$$

where TP, FP, and FN represent true positive, false positive and false negative rate.

$$F1 - score = \frac{2PR}{P + R} \tag{3.11}$$

where P represents precision and R represents recall.

We evaluate the segmentation task in terms of Dice Similarity Coefficient (DSC). DSC comes from the perspective of set theory and measures the similarity between two sets of data. Given two sets A and B, DSC is defined as follows

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3.12}$$

where $|A|$ and $|B|$ represents the cardinalities of the sets A and B. In our problem we use DSC to measure the spatial overlap between the predicted segmentation map and the ground truth label. To calculate the DSC score, we first binarize the predicted segmentation map with a threshold of 0.5 and then use the following

$$DSC(A, B) = \frac{2\sum_{i,j} a_{ij} b_{ij}}{\sum_{i,j} a^2{}_{ij} + b^2{}_{ij}} \tag{3.13}$$

where $A = (a_{ij})_{i,j=1}^{H,W}$ and $B = (b_{ij})_{i,j=1}^{H,W}$ represent the predicted output segmentation map and ground truth binary map respectively.

3.5.6 Performance Evaluation

In this section we assess the effectiveness of our proposed model in different settings. First we assess the effectiveness of multi-task learning on segmentation performance. Then we analyze the benefit of increasing data with classification label on network's performance. We further, assess the effect of masking technique and show the predicted result of our segmentation module on a randomly selected Fundus sample from our dataset.

	Disc Seg	**Cup Seg**
Methods	Dice	Dice
InterGD	95.7	**89.7**
InterGD-Seg	96.7	89.2
U-Net [26]	-	85.0
CNN [32]	**97.3**	87.1

TABLE I: Performance of InterGD and baselines for the segmentation task on DRISHTI-GS

	Prediction		
Methods	Precision	Recall	F-Score
InterGD	81.4	78.5	79.9
InterGD-Pred	81.0	84.0	82.5
CNN	85.3	86.3	85.8

TABLE II: Performance of InterGD and compared methods for the prediction task on I-ODA-B.

3.5.6.1 Effect of Multi-task Learning

We assess the effect of multi-task learning in addressing the shortage of segmented data by comparing our result with two previous approaches [26] and [32] for the disc-cup segmentation task introduced in Section 3.5.2. We also compare the result with the adopted version of our model InterGD-Seg. We employ DRISHTI-GS and I-ODA-B datasets for this experiment. Table I demonstrates the performance result on InterGD and the compared methods. The result for Sevastopolsky et [26] and Zilly et [32] are reported directly from their papers.

As Table I suggests, our proposed method outperforms the strongest baseline Zilly et [32] on cup segmentation by almost 2.6% and Sevastopolsky et by 4.7%. Moreover, it can be seen that

the adopted version of our model for single task of segmentation, InterGD-Seg, also surpasses the performance of Zilly et by 2.1% and Sevastopolsky et by 4.2% on cup segmentation without employing any data augmentation techniques or cropping the image to the region of interest as its employed in [26]. InterGD-Seg achieves a comparable result to our strongest baseline, Zilly et, on disc segmentation.

This result demonstrates the advantages of multi-task learning over single-task learning especially for more complex task such as cup segmentation. Table I shows that the segmentation task can benefit from the classification module that has access to adequate data with classification label while achieving clinical interpretability through the obtained disc and cup segmentation maps and CDR assessment.

Additionally, we evaluate an adopted version of our model to the single task of classification, named InterGD-Pred and compare it with a deep convolutional network CNN introduced in 3.5.2. The result is illustrated in Table II. The result in Table II shows that as expected CNN model that is specifically designed for the task of classification and trained solely for classification module with adequate training data performs better than InterGD which handles both tasks of segmentation and classification. Table II indicates that our interpretable model's performance is still comparable to an uninterpretable deep models such as CNN. The reason that our multi-task model InterGD does not outperform its single task version could be attributed to the difference between the distributions of the data employed in classification and segmentation tasks.

TABLE III: The effect of the classification training size on the performance of InterGD - employing DRISHTI-GS as the segmentation dataset.

	Disc Seg	**Cup Seg**	**Prediction**		
Pred Training Dataset (size)	Dice	Dice	Precision	Recall	F-Score
I-ODA-A ($\sim 8k$)	**96.4**	**90.3**	89.0	90.2	**89.6**
I-ODA-B ($\sim 2k$)	95.7	89.7	81.4	78.5	79.9

The overall result indicates that our proposed model: (1) alleviates the problem of shortage of segmented data with the help of the classification module, and (2) alleviates the lack of interpretability inherited in merely using a classification module by the obtaining disc and cup segmentation maps and CDR assessment. Furthermore, we showed that even with datasets following different data distributions, the segmentation component can still benefit from the data with only classification labels.

3.5.6.2 Effect of Classification Training Size

We analyze the effectiveness of our multi-task setting by employing a larger classification training set IODA-A containing almost $8k$ training samples which is almost 4 times larger than I-ODA-B. We perform this experimenst for both of the segmentation datasets, DRISHTI-GS and RIGA. The results are shown in Table III and Table IV. As the result suggests, increasing the amount of data for the classification task, boosts the performance of segmentation task

TABLE IV: The effect of the classification training size on the performance of InterGD - employing RIGA as the segmentation dataset.

	Disc Seg	**Cup Seg**	**Prediction**		
Pred Training Dataset (size)	Dice	Dice	Precision	Recall	F-Score
I-ODA-A ($\sim 8k$)	**96.1**	**83.0**	88.2	92.1	**90.1**
I-ODA-B ($\sim 8k$)	95.7	81.9	79.4	76.8	78.

further. This result indicates the effectiveness of multi-task learning and that segmentation task can benefit from the increase in data with only classification label. As classification labels are easier to obtain and more accessible, this result shows that the difficulty of procuring segmentation labels can be mitigated by exploiting the classification labels by formulating the problem in multi-task setting.

3.5.6.3 Effect of Masking

In this section, we assess the effectiveness of the masking technique by adopting two versions of our model, one with masking disabled and one with masking enabled. We employ both of the DRISHTI-GS and RIGA and I-ODA-B dataset for this experiment. The result for DRISHTI-GS and RIGA datasets are demonstrated in Table V and Table VI respectively. As Table V and Table VI suggest, the result shows a slight improvement in classification performance when masking is enabled. Although masking technique does not benefit the segmentation task, it

TABLE V: The effect of masking on the performance of InterGD on I-ODA-B and DRISHTI-GS datasets.

	Disc Seg	**Cup Seg**	**Prediction**		
Methods	Dice	Dice	Precision	Recall	F-Score
InterGD	95.7	89.7	81.4	78.5	**79.9**
InterGD (No Masking)	**96.2**	**90.0**	75.2	80.5	77.7

TABLE VI: The effect of masking on the performance of InterGD on I-ODA-B and RIGA datasets.

	Disc Seg	**Cup Seg**	**Prediction**		
Methods	Dice	Dice	Precision	Recall	F-Score
InterGD	95.7	81.9	79.4	76.8	**78.0**
InterGD (No Masking)	**95.9**	**82.1**	79.9	75.8	77.8

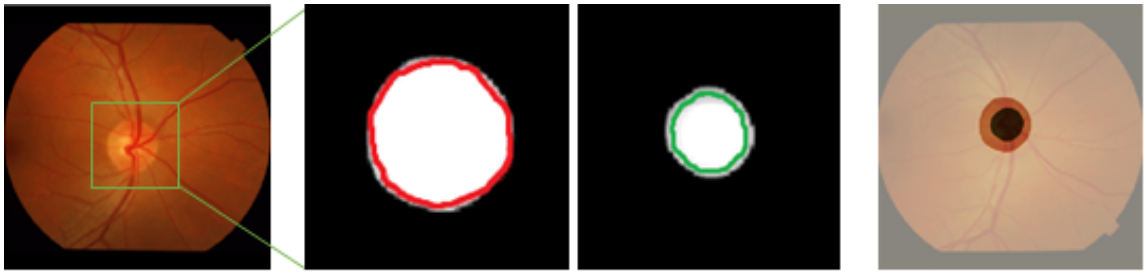


Figure 12: Illustration of predicted segmentation maps and ground truth on a sample fundus photo from DRISHTI-GS.

does not degrade its performance either. We can also see that the effect of masking technique lessens for both disc/cup segmentation and classification performance as the number of training samples for segmentation component increases.

3.5.6.4 Predicted Result Analysis

In this section, we visualize the segmentation maps predicted by InterGD on a random sample selected from DRISHTI-GS dataset. The result is demonstrated in Figure 12. In Figure 12 the full Fundus image on the left demonstrate the input and the Fundus with highlighted center area on the right illustrates predicted disc and cup regions. The middle images shows the zoomed in predicted and ground truth of disc and cup segmentation maps. The white regions are the predicted disc and cup. The red and green contours demonstrate the ground truth of disc and cup segmentation map respectively. As Figure 12 suggests, the proposed model achieves a reasonable accuracy in predicting the cup and disc segmentation maps.

3.6 Conclusion

In this chapter, we study the problem of clinical interpretability in the context of real-world ophthalmic imaging application, glaucoma detection from a full Fundus photo. Glaucoma is a complex disease that gradually damages the optic nerve and make its diagnosis very challenging. Cup to Disc ratio is commonly used in real-world clinical settings for assessment of optic nerve head and glaucoma screening. The current approaches mainly attempt at predicting glaucoma from a Fundud directly and disregard the clinical interpretability. The lack of interpretability pose a major challenge for clinical applicability of such methods.

To address this challenge, we propose a novel deep neural network that integrates segmentation and classification into a unified end-to-end architecture. We formulate our problem into a multi-task setting in which the segmentation component aims at localizing the area of optic disc and cup and targets the clinical interpretability of our model. Classification component predict glaucoma from fusion of disc and cup from segmentation module and alleviates the shortage of segmentation labels. We showcased the effectiveness of our framework through extensive experiments on several datasets. Our experimental results showed the superiority of our model over the previous works without the need for cropping the image to ROI or employing any data augmentation.

CHAPTER 4

SELF-SUPERVISED LEARNING FOR REAL-WORLD CLINICAL DATA APPLICATION

(This chapter was previously published as Mojab, N., Noroozi, V., Yi, D., Nallabothula, M. P., Aleem, A., Philip, S. Y., and Hallak, J. A.: **Real-world multi-domain data applications for generalizations to clinical settings**. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) , pages 677–684. IEEE, 2020 [3].)

4.1 Introduction

Deep learning methods have proved to excel in many fields of computer vision. Motivated by these successes, deep learning application in medical imaging, in particular ophthalmic imaging applications, have received considerable attention over the past decade. This has inspired an exponential growth of many research studies aiming to address various ophthalmic imaging problem using deep learning methods from disease classification to lesion segmentation or progression prediction. Most of the current research studies focus on solving the problem by relying on standardized data from artificial settings, such as clinical trials. However, in many real-world clinical applications, data is very complex and lacks standardization. Therefore, the standardized data do not necessarily represent the characteristics of real-world clinical data yielding varying results in clinical translation.

Real-world data is characterized by variability in quality, machine-type, setting, source, lacks standardization and labels, and embedded with inherent noise sourcing from human errors. The complexity of real-world data makes the translation of deep learning model to real-world applications very challenging. Hence, there has been a growing concern about the generalization capacity of deep learning based models for real-world clinical data applications and hence their translation to real-world clinical settings. Although, the deep learning based models have been extensively studied in medical imaging, there has been very limited work on the generalization and translation capacity of such methods to real-world clinical settings. Different from other fields of computer vision, collecting medical data that represent the characteristic of real-world data is an ongoing challenge. Acquiring labeled data is not only laborious and expensive, the access to data is limited due to HIPAA regulations, patient privacy, data ownership, and challenges regarding data collection from different sources. These challenges results in shortage of annotated datasets in medical field, in particular medical imaging. Consequently, most of the medical imaging problems with limited data are formulated in transfer learning setting, relying on pretrained networks on non-medical data such as ImageNet [17] or CIFAR [36] for feature extraction. These methods tend to perform to their full potential when the nature of data in the downstream task is similar to the data used for pretraining networks. However, the nature of medical data is quite different from imaging data in datasets such as ImageNet or CIFAR which limits the capacity of these methods for extracting effective visual representations for medical data.

In general, acquiring unlabeled data is easier and more accessible compared to labeled data. Self-supervised learning methods have proved to excel in learning representations by exploiting the unlabeled data [37]. Moreover, the recent research studies have shown that self-supervised learning can learn more generalizable features and hence lead to better generalization over supervised methods [38].

Motivated by these challenges, we develop a framework, called FundusNet, that exploits unlabeled data by employing a self-supervised learning aiming to learn effective visual representation for ophthalmic imaging data, in particular Fundus photos. This builds on the work of [38] where we incorporate a collection of data augmentations into a contrastive learning framework to improve the generality of learned representations. Further, we assess the generalization capacity of deep learning algorithms and their translation to real-world data applications by exploring their capacity in coping with complexity of real-world data versus standardized data. In order to evaluate our proposed approach, we perform comprehensive experiments on a real-world clinical application, glaucoma detection, using a collection of datasets including a dataset that represent real-world multi-domain and a dataset representing standardized single-domain data. We demonstrate the effectiveness of self-supervised representation learning by comparing our work against fully supervised methods. We also show the quality of the learned representation for Fundus photos by comparing our framework to pretrained networks using non-medical datasets. To the best of our knowledge, this chapter is the first work aiming to build a network for medical imaging representation learning via self-supervised learning and as-

sess the clinical translation and generalization capacity of deep learning methods on real-world multi-domain data.

The rest of this chapter is organized as follows. We start by a brief overview on related works on self-supervised learning for visual representation learning in Section 4.2. Then we introduce our approach and problem formulation in Section 4.3. Section 4.4 reports the experimental setting and performance results. In Section 4.5 we conclude the chapter.

4.2 Related Works

The machine learning field has witnessed an evolution of self-supervised learning methods over the past decades. Some of the earliest studies in the literature target the problem of self-supervised learning by relying on specific design choices for the network architecture or predictive pretext tasks [39–42]. These limitations, could potentially restrict the generality of the learned representation and hence limit the applications of such methods to medical imaging that deals with complex real-world multi-domain datasets. So it is desired to learn more generalizable visual representations that can be utilized across different domains. The self-supervised visual representation learning proposed in [38], however, avoids the limitation of specific design choice or complexity of solving predictive tasks by incorporating a broad family of augmentations and contrastive loss into a simple off the shelf architecture. Inspired by the work in [38], we leverages a collection of augmentations suited for our medical dataset embedded in a simple network architecture. Incorporation of data augmentation into a contrastive learning framework could potentially lessen the sensitivity of the model to domain-specific information in

the data and hence improves upon the generality of learned representations and generalization capacity of the model.

4.3 Methodology

Self-supervised representation learning is the task of learning features without the human supervision, in particular without the requirement of human input in terms of labeling. In our work, we first learn the visual representation of data via a self-supervised learning method and then show the quality of learned representation in the context of a downstream task, glaucoma detection.

4.3.1 Contrastive Learning

In the constrastive learning framework, the representations are learned by maximizing the similarity between the two views of the same data input. The framework consists of four main components as follows

- Augmented views: Given the input x_k, we obtain two augmented views of x_k denoted by $x_k^{(1)}$ and $x_k^{(2)}$, by applying stochastic data augmentations.
- Base encoder: The base encoder $f(.)$ maps the augmented inputs $x_k^{(1)}$ and $x_k^{(2)}$ to the feature maps $z_k^{(1)}$ and $z_k^{(2)}$. The encoder function is modeled by a neural network.
- Projection head: A projection head $q(.)$ is responsible for mapping the feature maps $z_k^{(1)}$ and $z_k^{(2)}$ to $h_k^{(1)}$ and $h_k^{(2)}$ in the latent space to which a contrastive loss is further applied. The function $q(.)$ is also modeled by a neural network.
- Contrastive prediction task and constrastive loss: In the context of this problem, two augmented views of the same input are considered the positive pair, i.e. $x_k^{(1)}$ and $x_k^{(2)}$.

The contrastive prediction task aims to identify $x_k^{(2)}$ for a given $x_k^{(1)}$ via a contrastive loss function.

Given the Fundus photo in our dataset, the application of two augmentations results in two augmented views of the input Fundus which are further being mapped to representation z via the encoder function f that are further used in downstream tasks. The overall architecture of the FundusNet framework is illustrated in Figure 13.

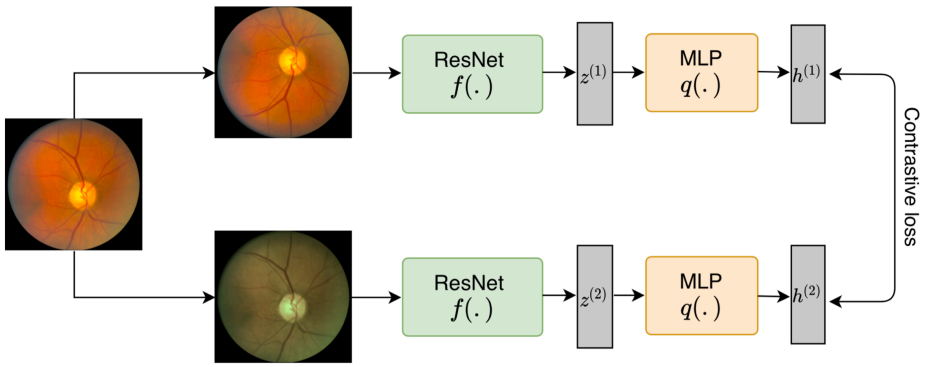


Figure 13: The overall architecture of contrastive learning framework for FundusNet.

Given a batch of N samples, applying two augmentations results in $2N$ samples. Given each positive pair, the remaining samples in the batch $2(N-1)$ are considered negative samples. Given a positive pair $\{(1),(2)\}$, the contrastive loss function is defined as follows

$$\ell_{(1),(2)} = -log\frac{exp(sim_{(1),(2)})/\tau}{\sum_{i=1}^{2N} \mathbb{1}_{[(i)\neq(1)]} exp(sim_{(1),(i)})/\tau} \tag{4.1}$$

where τ represents a temperature parameter, $\mathbb{1}$ is an indicator function, and $sim_{(i),(j)}$ represents the cosine similarity. The cosine similarity between two vectors a and b is defined as follows

$$sim(a,b) = \frac{a^T b}{\|a\|\|b\|} \tag{4.2}$$

The overall loss function is defined as the mean of losses for all positive pairs as follows

$$\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{N}[\ell((2i-1),(2i)) + \ell((2i),(2i-1))] \tag{4.3}$$

4.3.2 Classification Problem

Let $\mathcal{D} = \left\{(x_k, y_k)|y_k \in \{0,1\}\right\}_{k=1}^{M}$ denotes the training set for our downstream task of glaucoma detection. M represents the number of samples, and y_k indicates the binary label of input x_k where values of 1 and 0 represent diseased and non-diseased class respectively. Given $\mathcal{D}$, our goal is to learn a binary classifier function $f_c : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ_c. We define the following functions

$$f(x_k;\theta_f) = z_k \tag{4.4}$$

$$g(z_k; \theta_g) = y_k \tag{4.5}$$

where $f(.)$ parameterized by θ_f represents the encoder function $f : \mathcal{X} \to \mathcal{H}$ introduced in the previous section and it is learned by unsupervised pretraining using the unlabeled Fundus data. The function $g(.)$ that is parameterized by θ_g represents the decoder function $g : \mathcal{H} \to \mathcal{Y}$ mapping the learned representation z_k to the label space. Given the input $x_k \in \mathcal{X}$, function f_c can be decomposed such that

$$f_c(x_k, \theta_c) = (g \circ f)(x_k) \tag{4.6}$$

where $\theta_c = \{\theta_g, \theta_f\}$. Given the input x_k, binary classifier $f_c(.)$ estimates the probability of an input image being diseased.

Given the training set $\mathcal{D}$, the loss function over the M samples is defined as follows

$$\mathcal{L}_c(\mathcal{D}; \theta_c) = \sum_{k=1}^{M} \ell_c(x_k; \theta_c) \tag{4.7}$$

where $\ell_c(x_k)$ represents the classification loss for sample x_k and it is defined as cross-entropy between the model's estimation and the ground-truth label

$$\ell_c(x_k; \Theta_c) = -y_k \log \hat{y}_k - (1 - y_k) \log(1 - \hat{y}_k)) \tag{4.8}$$

where $\hat{y}_k = f_c(x_k; \Theta_c), 0 < \hat{y}_k < 1$ indicates the model's prediction for input x_k and y_i represent its corresponding ground-truth label.

4.4 Experiments

4.4.1 Data Collection

We employ I-ODA dataset [1] for our experiments. We isolated 6244 glaucoma and 7664 non-glaucoma patients from this dataset. Non-glaucoma patients in our dataset are selected as those patients being diagnosed with neither glaucoma, nor glaucoma suspect with no potential damage to the optic nerve head. Among the 12 categories, we isolated Fundus images for the purpose of our experiments in this work which we refer to as ODA-G dataset. Each Fundus image can be generated by a different imaging device. The statistics of image data and the corresponding imaging device distribution in the ODA-G dataset is illustrated in Figure 14. The ODA-G dataset consists of image data from 11 different imaging devices where devices A, B, C, and D comprise 96% of the data, while the other 7 devices are responsible for almost 4% of the data. A snapshot of samples of Fundus photos from 4 major imaging devices A, B, C, and D are illustrated in Figure 15. As the Figure 15 suggests, images across different devices can have different color distributions, different shapes, and position of the cup and disc regions. We regard the imaging devices in I-ODA dataset as different domains. Therefore, the ODA-G data is a multi-domain dataset that is a mixture of data generated from different imaging devices. We employ ODA-G for unsupervised pretraining which learns the base encoder network (function f) without labels.

In order to assess the deep learning algorithms in coping with real-world data and their generalization capacity as opposed to simpler standardized data, we isolated two subsets of data as follows

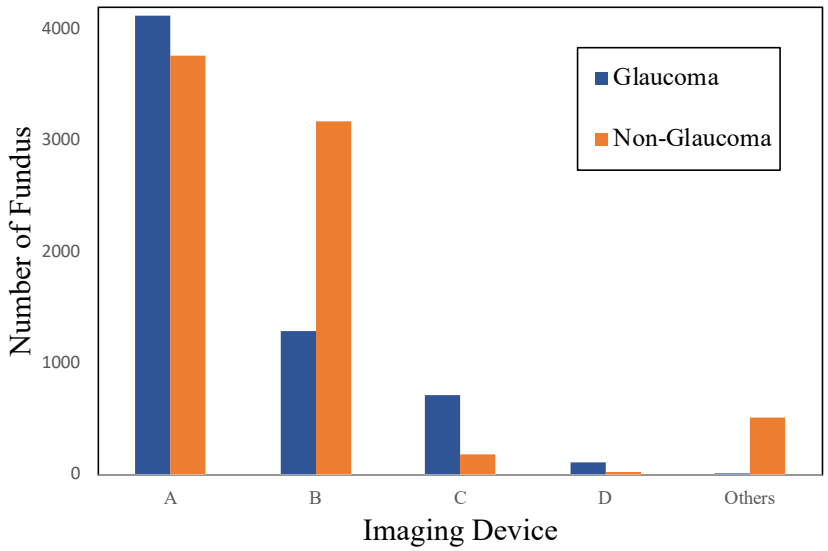
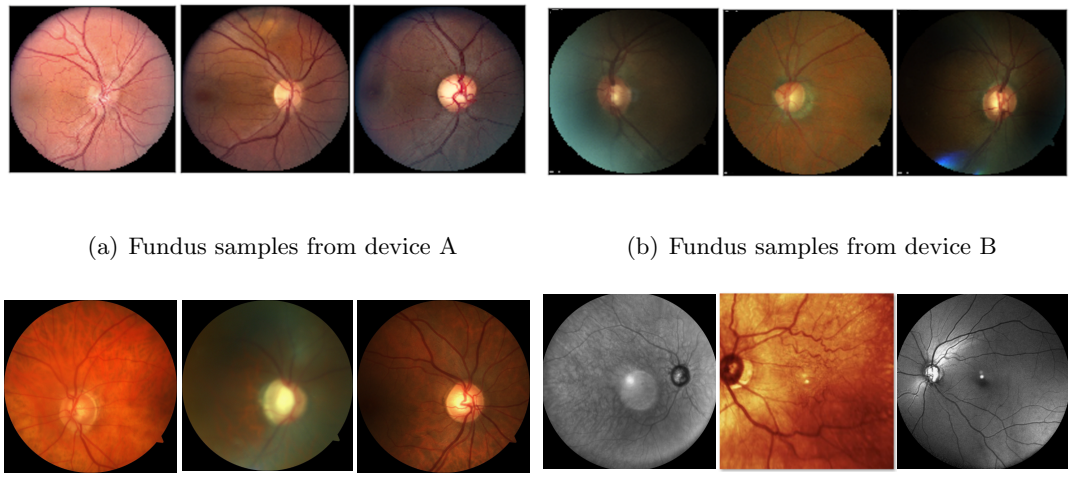


Figure 14: The number of Fundus photos per imaging device.

- Fundus images generated from only one imaging device, device A, which we refer to as ODA-A. Although this dataset, is also a subset of a real-world dataset, we regard it as standardized dataset as it contains data from a single domain Hence it resembles the characteristic of standardized dataset from artificial settings that are usually single domain.
- ODA-G that contains Fundus generated by all imaging devices used for glaucoma screening in our dataset. This dataset represents a multi-domain data capturing the complex aspect of diverse real-world data.



(a) Fundus samples from device A

(b) Fundus samples from device B

(c) Fundus samples from device C

(d) Fundus samples from device D

Figure 15: A snapshot of samples of Fundus photos from I-ODA dataset generated by different imaging devices.

We aim to answer two main research questions through our experiments, (1) how deep learning models cope with the complexity of real-world data comprised of multiple domains versus standardized datasets with a single domain., and (2) how important training the model on real-world data is for generalizations to a clinical setting.

4.4.2 Baselines

To show the effectiveness of our approach, we compare our result against fully supervised methods and pretrained nwtorks on ImageNet and CIFAR. Previously proposed works on glau-

coma detection, mainly employ a Convolutional Neural Network (CNN) using one of the standardized datasets. We simulate the general approach and adopt two supervised methods.

- We employ ResNet-50 with width of 1 as the base encoder network. The decoder is chosen as a simple MLP that follows the structure of 1 hidden layer of size 1024, batch normalization, ReLU activation, and dropout. Further, a sigmoid activation function is applied to the obtained output.
- A convolutional neural network (CNN) that is specifically designed for our task and dataset. This network is composed of 2 convolutional block, where each consists of 1 convolutional layer with ReLU activation, batch normalization, non-linearity ReLU, and max pooling. Two fully connected layers with size 512 and 64 followed by batch normalization, non-linearity ReLU, and dropout are further applied. The sigmoid activation function is used on the final layer to predict the output.

We also compare FundusNet with two other pretrained networks from [38] on ImageNet and CIFAR-10 dataset in a transfer learning setting. Based on the choice of CIFAR-10 based encoder or ImageNet based encoder, a slightly different network architecture are employed. For ImageNet based encoder, a ResNet with depth 50 and three hidden layer width choices ($1\times$, $2\times$, $4\times$) are employed [38], [43]. For CIFAR-10 based encoder, a simpler ResNet like network with depth of 18, and width 1 is employed.

4.4.3 Experimental Setting

4.4.3.1 Network Architecture

The base encoder f and projection function q are both modeled with neural networks. Similar to [38], we take the base encoder to be ResNet architecture with depth of 18 and width set to 1. The projection head is taken to be a 2 layer MLP. The overall base encoder network consists of four main residual blocks each consists of two stacked convolutional followed by batch normalization.

4.4.3.2 Training and Optimization

We employ slightly different experimental setup based on the choice of pretrained networks. For FundusNet, images were resized to 112×112. Resize with random flip and color distortions are employed for data augmentation. For ImageNet based pretrained network, images were resized to 224×224. Data augmentation are chosen as random crop and resize with random flip, color distortions, and Gaussian blur [38]. For CIFAR-10 based pretrained network, images were resized to 32×32. Random crop and resize with random flip and color distortions are employed for data augmentation [38].

We train the supervised methods with and without data augmentation. When data augmentation is used, color distortion and random flip are employed.

The split of training and testing for our dataset is based on the patient ids. The reason for that is that in our dataset (ODA-G and ODA-A), each patient can have more than one Fundus photo taken coming from different exam sessions. Therefore, to prevent the leakage of any information from the test set into training, we split the data by patients. We chose

a split of 80% and 20% for training and testing respectively. The hyperparameter tuning for each dataset is performed on the validation set chosen randomly as 20% subset of the training set. After finding the appropriate values for hyperparameters, all the training set is used for training. Eventually, the result on the test data is reported in terms of the accuracy metric. For pretraining, we use LARS optimzer [44] and we use SGD and Adam for downstream classification tasks. For SGD optimizer with Nestrov momentum the momentum parameter was set to 0.9 and the initial learning rate is selcetd from $\{0.1, 0.07, 0.01, 0.007, 0.001\}$. Batch size is selected from $\{64, 128, 256, 512, 1024\}$

4.4.3.3 Transfer Learning

We evaluate the transfer learning setting across two datasets ODA-G and ODA-A by (1) learning a linear classifier on top of the pretrained network using FundusNet (ODA-G), ResNet (CIFAR, ImageNet) based encoder and (2) fine-tuning $x\%$ of the network where $x \in \{25, 50, 75, 100\}$ based encoder network. In the first setting, a logistic regression classifier is trained on top of the frozen base encoder network. The extracted features from the pretrained network are used to perform binary classification for detecting glaucoma in a given Fundus input.

In the first setting, no data augmentation is applied. The number of epochs is selected from $\{50, 100, 200, 300, 500\}$, batch size is selected from $\{1024, 512, 256, 128\}$ and learning rate is selected from $\{0.1, 0.07, 0.05, 0.03, 0.01, 0.007, 0.001\}$. In the second setting, we fine-tune the base network using the pretrained network's weights as initialization. We fine-tune $\{25\%, 50\%, 75\%, 100\%\}$ of the network for $\{50, 100, 200, 300, 500\}$ epochs where learning rate is selected from $\{0.1, 0.07, 0.05, 0.03, 0.0$

The batch size is selected from $\{512, 256, 128, 64\}$ and the rest of hyperparameters are set to default value as reported in [38].

4.4.4 Experimental Results

In this section we demonstrate the effectiveness of our approach by comparing against fully supervised models and transfer learning setting using ImageNet and CIFAR based pretrained network. Further, we answer our two research questions in Section 4.4.1 through extensive experiments on real-world clinical application, glaucoma detection, using both real-world data and standardized dataset.

4.4.4.1 Comparison Against Fully Supervised Approaches

In this section, we demonstrate the effectiveness of FundusNet in learning visual representation of Fundus photos by comparing to pretrained networks on ImageNet, CIFAR and fully supervised models. We show the result of our work on both ODA-G and ODA-A dataset for the task of glaucoma detection. We formulate the problem in transfer learning framework and evaluate the model in both settings, linear evaluation and fine-tuning.

In the first setting, we evaluate our approach by learning a linear classifier using self-supervised representations learned by pretrained base encoder network and compare the result against fully supervised methods. In the second setting, we explore the performance of our classifier with fine-tuning the base encoder network using the weights of the pretrained network as initialization. We explore fine-tuning $\{25\%, 50\%, 75\%, 100\%\}$ of the network. The result is demonstrated in Table VII.

TABLE VII: Comparison of employing self-supervised learned representations via transfer learning (TR) in linear evaluation and fine-tuning settings against fully supervised baselines.

	ODA-G	ODA-A	Weights
Supervised Baselines			
ResNet-50 (1×)	59.93	83.57	Random
ResNet-50 (1×) (+Augmentation)	82.61	86.44	Random
CNN	64.67	88.35	Random
CNN (+Augmentation)	82.93	90.43	Random
self-supervised TR+Linear			
ResNet	72.87	82.40	CIFAR-10
ResNet-50 (1×)	80.26	88.92	ImageNet
ResNet-50 (2×)	83.84	91.00	ImageNet
FundusNet	**84.87**	**91.41**	ODA-G
self-supervised TR+Fine-tuning			
ResNet	82.05	86.04	CIFAR-10
ResNet-50 (1×)	83.28	90.35	ImageNet
ResNet-50 (2×)	83.14	90.43	ImageNet
FundusNet	**84.96**	**91.88**	ODA-G

The result in Table VII shows the superiority of FundusNet using self-supervised learned representations over fully supervised approaches. We can also see that using FundusNet for feature extraction outperforms the pretrained networks on ImageNet and CIFAR. We can achieve superior result over ResNet-50 (1×) and ResNet-50 (2×) with much simpler network, smaller image size for inputs and fewer number of parameters. The overall result shows that FundusNet is able to learn more generalizable features for medical images, Fundus and hence lead to better generalization for medical imaging downstream tasks.

Among the supervised baselines in Table VII, the CNN (+Augmentation) method that is specifically designed for our dataset-task achieves a better result over off the shelf ResNet networks. This result suggests that specifically designed networks for each dataset-task combination are usually important to the success of supervised approaches especially when data is limited. This limitation could potentially limits their applications and generalization capacity, especially in real-world clinical setting where the data is complex and multi-domain.

The superiority of our framework over supervised baselines shows that we can avoid the complexity of design choice for each particular task by simply using one of the off the shelf networks. This result is particularly important for medical imaging where data is limited to train robust supervised models.

4.4.4.2 Neural Networks On Real-world Data

In this section, we assess the capacity of neural networks in coping with real-world datasets versus standardized datasets by answering our two research questions. How well deep learning

based models cope with the complexity of real-world data. What the role of real-world data is in generalization and translation to clinical settings.

We answer the first question by analyzing the results of our work on the two datasets ODA-G and ODA-A. The ODA-G dataset consists of data from multiple imaging devices forming a complex multi-domain data representing the characteristics of real-world data. The ODA-A dataset comprises a single domain and represents the characteristics of a simpler and more standardized dataset that are commonly used in current approaches in the literature for glaucoma detection.

The result from Table VII shows, when using ODA-A as training set, the model achieves superior result over using ODA-G as the training set in all three settings, supervised, linear evaluation, and fine-tuning. The performance gap is particularly noticeable when using supervised learning methods. The reason is that the success of fully supervised methods usually relies on the availability of large standardized labeled datasets. This experiment verifies that deep learning models do not perform as well on complex real-world datasets as on standardized datasets and their performance degrades as the complexity of data increases.

To answer the second question, we design an experiment where we form four pairs of (ODA-G, ODA-G), (ODA-G, ODA-A), (ODA-A, ODA-G) and (ODA-A, ODA-A) datasets. We aim to investigate how training the model with either real-world data or standardized data affects the generalization capacity of the model to the data from the same or other settings and domains. We train the model on the first element of each dataset pair and perform evaluation on the second element of a dataset pair. The result is illustrated in Table VIII.

TABLE VIII: Generalization across pairs of datasets employing self-supervised learned representation via transfer learning.

Training Data	Testing Data	Test Accuracy(%)
ODA-G	ODA-G	84.87
ODA-G	ODA-A	**90.35**
ODA-A	ODA-G	**76.53**
ODA-A	ODA-A	91.41

As Table VIII suggests, if we only train the model on one small unique dataset, ODA-A, and perform the evaluation on an extremely complex real-world dataset, ODA-G, we achieve the worst result depicted in red in Table VIII. On the other hand, when we train the model using a diverse multi-domain real-world dataset, ODA-G and evaluate it on a smaller and more unique dataset, ODA-A, we achieve a promising result as depicted in green in Table VIII. This result indicates that even though complex data makes the learning process harder, it leads the model towards learning more generalizable features. It is not surprising that (ODA-A, ODA-A) experiment achieves the best result, as this is where usually neural networks perform to their full potential. However, we can see that the obtained result from the (ODA-G, ODA-A) experiment is also competitive with the result of the (ODA-A, ODA-A) experiment, indicating the advantage of learning with real-world data.

TABLE IX: Generalization across pairs of datasets employing supervised approaches.

Training Data	Testing data	Test Accuracy (%)
ODA-G	ODA-G	82.93
ODA-G	ODA-A	**74.00**
ODA-A	ODA-G	**69.19**
ODA-A	ODA-A	90.43

The overall results validates the crucial role of real-world data for generalizition and translation to clinical settings. If the model trains on small unique datasets, it fails to generalize to other settings and domains. Hence we need diverse datasets that capture the aspects of real-world data to cope with generalizations to other domains, and improve the translatable capacity especially in clinical settings.

Further we investigate the generalization capacity of self-supervised approaches in comparison with fully supervised methods in the similar setting. We perform the same experiments but under supervised settings. The result is shown in Table IX. The observation from Table IX (shown in bold) supports the previously obtained result from Table VIII. Training with a real-world dataset, ODA-G, generalizes better to ODA-A compared to the (ODA-A, ODA-G) experiment where we train the model on ODA-A and evaluate it on ODA-G. However, we can see that the performance gap between (ODA-G, ODA-G) and (ODA-G, ODA-A) experiment

and between (ODA-A, ODA-A) and (ODA-A, ODA-G) is more noticeable compared to the same experiments under sel-supervised setting in Table VIII. This result could indicate that supervised approaches perform to their full potential when they are trained and evaluated on the datasets from the same domain. Moreover, the comparison between the results in Table VIII and Table IX shows that (ODA-G, ODA-A) experiment performs poorly under a supervised approach achieving only 74% accuracy, while under a self-supervised approach it achieves 90.35% as shown in Table VIII, which is more than 16% improvement over the supervised setting. This result verifies the superior capacity of self-supervised approaches in generalization to different device and domains over supervised approaches.

4.4.5 Performance Analysis

In this section, we first analyze the effect of data augmentation on generalization performance. Then, we evaluate the effect of training time and fine-tuning $x\%$ of the network on the performance of the model.

4.4.5.1 Effect Of Data Augmentation On Generalization

The work in [38] shows that incorporating a broad range of data augmentation in the self-supervised learning framework enhance the generality of learned representations. In this section we analyze the role of data augmentation in generalization for fully supervised approaches in both the presence and absence of data augmentation. For ODA-G we applied the composition of color distortion and random flip and for ODA-A we only applied the random flip. As the result in Table VII suggests, data augmentation can efficiently improve the generalization performance. This result shows that incorporating data augmentation enhances the capacity of the model

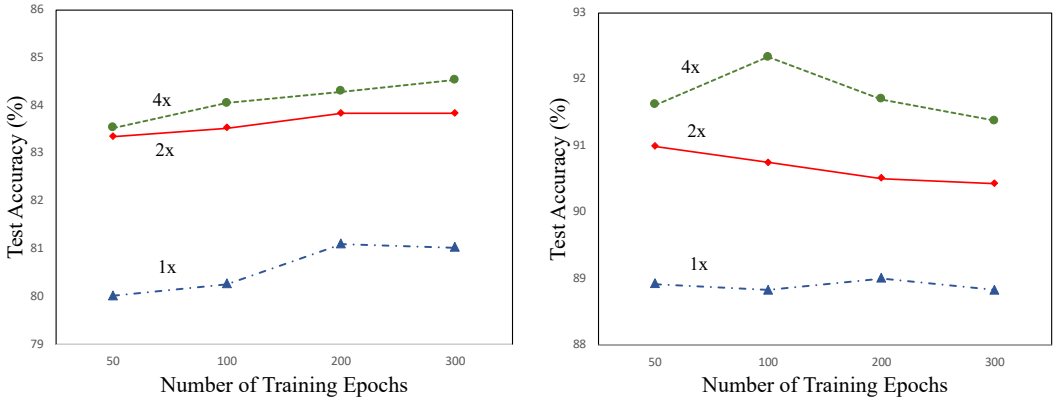
to learn more generalizable features. This improvement is particularly more noticeable for the ODA-G dataset that is comprised of multiple domain. Therefore, data augmentation can benefit the model training on a diverse multi-domain real-world dataset and improve upon their generalization capacity.

4.4.5.2 Effect Of Training Time On Real-world Data

In this section, we assess the effect of training time on the performance of glaucoma detection using pretrained networks on ImageNet and CIFAR by increasing the number of training epochs while keeping the batch size fixed for each model. Figure 16 shows the plots of test accuracy versus training epochs for linear evaluation employing ImageNet based encoder (ResNet-50 ($1\times$, $2\times$, $4\times$)).

Figure 17 shows the plots of test accuracy versus training epochs for linear evaluation employing CIFAR-10 based encoder (ResNet with depth 18 and width 1) on both ODA-G and ODA-A datasets.

As the plots in Figure 16 suggest, the performance improves as we increase the number of training epochs when employing the ImageNet based encoder network on the ODA-G dataset. However, increasing the number of epochs has the opposite effect on the ODA-A dataset when using ImageNet based encoder network. As Figure 17 suggests, we observe a similar behaviour when using CIFAR-10 based encoder. However, we can see that increasing the number of epochs has less effect on performance improvement on ODA-G compared to using the ImageNet based encoder. The overall result indicates that more diverse datasets with large amount of data can benefit more from longer training.



(a) ODA-G dataset

(b) ODA-A dataset

Figure 16: Effect of the training time on test accuracy employing ImageNet based pretrained network for ODA-G and ODA-A datasets.

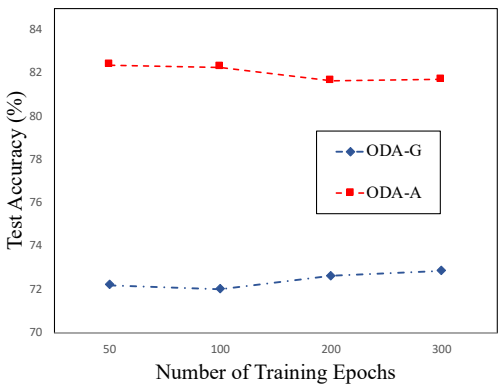


Figure 17: Effect of training time on test accuracy using CIFAR-10 based pretrained network for ODA-G and ODA-A datasets.

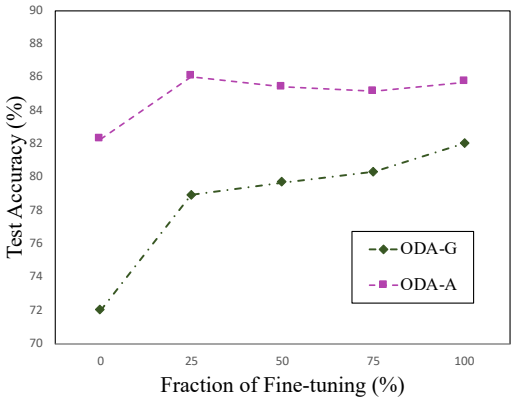


Figure 18: Effect of fine-tuning x% of the network on test accuracy using CIFAR-10 based encoder network for ODA-G and ODA-A.

4.4.5.3 Effect Of Fine-tuning percentage

In this section, we assess the effect of percentage of fine-tuning on performnace of the model. We plot the behaviour of the model when fine-tuning $x\% \in \{25, 50, 75, 100\}$ of the network on test accuracy and employing only the CIFAR-10 based encoder network. The result is shown in Figure 18. As the plot in Figure 18 shows, the test accuracy improves as we fine-tune a larger portion of the network when training on ODA-G dataset. Training on the ODA-A dataset still benefits from fine-tuning the network especially when it is fine-tuned on 1/4 of the network but after that, we do not observe any significant change in the performance. We achieved our best result for this experiment by fine-tuning the whole network from scratch when using the ODA-G dataset and 25% of the network when using ODA-A datset. We also observed that the model does not benefit substantially from fine-tuning the network with ImageNet weights.

4.5 Conclusion

In this chapter, we study the problem of generalization to real-world clinical settings when training deep learning models using real-world data versus standardized datasets. Real-world data is characterized by variability in quality, machine type, settings and lacks standardozation. The Complexity of real-world data pose a major on training deep learning models and their generalization performance. Most of the existing approaches in medical imaging applications, rely on standardized datasets collected from artificial settings such as clinical trials for modeling and problem solving which has raised a growing concern on their translation applicability in real-world clinical settings. Moreover, due to limited data in many medical applications, these problems are commonly formulated in transfer learning setting using pretrained networks on

data such as ImageNet. The dissimilarity of medical data and ImageNet like data, limits the capacity of such networks to extract effective representations for medical imaging data.

Motivated by this challenge, in this chapter, we presented a feature extraction network, called FundusNet, for ophthalmic imaging applications using self-supervised visual representation learning. We also assessed the generalization and translation capacity of deep learning algorithms by formulating the problem of real-world clinical application in a transfer learning setting using the learned representation from FundusNet. We showed the effectiveness of our approach through extensive experiments in the context of real-world clinical application, glaucoma detection. We showcased the importance of learning with real-world data for generalization, through performing extensive experiments on a multi-domain real-world dataset versus a single-domain standardized dataset. We showed that without learning on complex real-world data, the deep learning models cannot generalize well to clinical settings.

CHAPTER 5

CLASSIFICATION FOR SMALL DATASETS

5.1 Introduction

Contemporary deep neural networks relies heavily on large amount of training data to perform to their full potential in modeling and solving problems. For this reason, large scale datasets have been collected [17, 45, 46], enabling the development of powerful models pushing the state-of-the-art further in many downstream tasks of computer vision. The generalization capacity of deep neural networks tend to degrade when the data is scarce. This problem can be more sever for classification networks modeled by deep neural networks. However, there are several fields such as medical field that are facing many applications where access to large amount of data is very challenging or infeasible, e.g. rare diseases. The data collection problem can be even more challenging in domains such as healthcare where even the access to data is limited due to patient privacy regulations and data ownership.

Transfer learning and multi-task learning methods have proved to be an effective solution to the applications with limited data. These methods exploit the knowledge from other related tasks to improve upon the generalization performance of the model. Transfer learning improves the performance by exploiting the weights of a pre-trained network as initialization [47]. Despite their promising result their applications remains limited. The reason is that transfer learning relies on assumption of large amount of data for pre-training. These approaches also

are expected to performs to their full potential when the nature of data for the target task is similar to the one used for the pre-trained networks.

Multi-task learning methods jointly train the network with a group of related tasks to improve upon the generalization in all tasks [48, 49]. Classification tasks are commonly being accompanied by segmentation task into a multi-task framework [2]. Although, multi-task learning could be a potential solution for applications with limited data, they face the challenge of loss balancing in different heads which itself could lead to overfitting problem. Moreover, obtaining segmentation labels for dataset is not only challenging but also requires the domain knowledge in many fields such as helathcare.

Empirically speaking, segmentation methods have shown more robustness to overfitting when trained with small dataset [31, 50]. The reason could be potentially due to encoding a dense pixel-wise loss that incorporates a high-bias shape prior into the learning process. In this chapter, we introduce a novel framework to the above problem, called CvS (Classification via Segmentation). CvS harnesses the power of segmentation to learn from small datasets enabling us to perform classification for extremely small data ($\sim 1-5$ samples per class). CvS is single headed approach to multi-task learning that eliminates the need to balance the losses from different heads while still doing both tasks of segmentation and classification. CvS, alleviates the difficulty of procuring segmentation labels for a dataset by employing a label propagation method that obtain fully segmented data by segmenting only a small subset of the dataset. In order to evaluate our model, we perform extensive experiments on diverse problems and show the effectiveness of our proposed approach by achieving much higher classification result

when only a handful of data points is available to the model. To the best of our knowledge, this chapter is the first work aiming to address classification for extremely small datasets by utilizing segmentation.

The rest of this chapter is organized as follows. We start by brief overview of related works on image classification, particularly in the small data regime in Section 5.2. In Section 5.3 and Section 5.3.1 we present our model and some preliminary concepts on the problem. In Section 5.3.2 we explain how we incorporate two simple approaches to obtain segmentation label for a dataset. Section 5.5 presents the experimental results on diverse image classification problems using four datasets. In Section 5.6 we conclude the chapter.

5.2 Related Works

The machine learning field has witnessed an evolution of deep classifier networks over the past decades developing from a simple multi-layer network [51] to more complex and deeper neural networks aiming at pushing the state-of-the-art in image classification problem [52–57]. The current state-of-the-art models that are proposed for image recognition task, however, are mainly developed based on the assumption of availability of large amount of training data [58–62].

There are several studies that aim to tackle the problem of small training set by formulating the problem into a transfer learning setting or multi-task learning. Transfer learning, exploits the knowledge from pre-trained networks by utilizing the weights of pre-trained network as initialization [63, 64]. These methods have shown promising results when performed on the downstream tasks with the nature of data similar to the data used for transfer learning. How-

ever, very little works has been done on applications with datasets that are very different in nature from the data used in pre-trained networks. Additionally, pre-training large networks on large datasets requires powerful computational resources that may not be available to everyone.

Multi-task learning have proved to be an effective way to improve upon the genralizzation performance by jointly training the network with group of related tasks. One of the most commonly studied multi-task setting is grouping the classification with segmentation task [2]. However, the loss balancing from different head in mult-task learning methods, often lead to overfitting problem. Moreover, acquiring segmented data is very time-consuming and laborious limiting the practical applications of such approaches. Some studies aimed to address the problem by incorporating the power of transfer learning into a multi-task learning framework to exploit the benefits from both settings [65].

Our approach is relevant to the multi-task learning but there are significant differences between them. Our method derive classification result from segmentation module, forcing both tasks to be computed together. Therefore, our approach is single headed and eliminates the need for balancing the loss in different heads. This allows for approach to achieve a much higher classification result when only a handful of example is available for training.

5.3 Methodology

CvS is a single-headed approach to the standard multi-task learning methods, which eliminates the need to balance the losses from different heads. Our proposed method derives the classification result via a segmentation module forcing both tasks, segmentation and classification, to be computed together. The overall framework of CvS and its comparison to standard

classifier networks and multi-task learning is illustrated in Figure 19. Empirically, segmentation networks have proved to be more robust to overfitting when trained with small datasets. CvS harnesses the power of segmentation to learn from smaller datasets by encoding a dense-loss and incorporating a high-bias shape prior into learning process. This allows us to perform classification on extremely small datasets ($\sim 1-5$ samples per class) and achieve higher classification result in a low data regime.

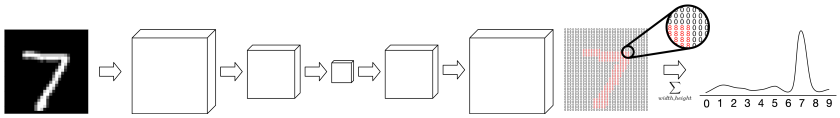
5.3.1 Problem Formulation

Before presenting the image classification via segmentation, we first introduce the notations that will be used throughout this chapter. Image Classification, is the task of automatically classifying an image into a subset of predefined classes. Segmentation, is the task of automatically classifying each pixel of an image into a predefined class. Let $\mathcal{D}$ denote the entire training set, which consists of N samples and defined as

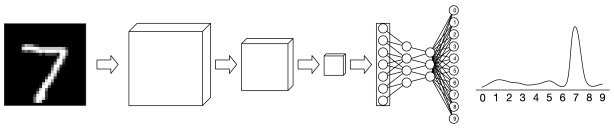
$$\mathcal{D} = \left\{(x_k, s_k, y_k) | s_k(i,j) \in \{0,1\}, y_k \in \{1,2,...,P\}\right\}_{k=1}^{N}$$

where x_k is of size $[H, W, C]$ with H, W, C representing the height, width and number of channel, s_k represents the segmentation map and y_k represents the ground truth class label with P representing the number of classes. Given $\mathcal{D}$, our goal is to learn a classifier $f_c : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ_c. We define the following functions.

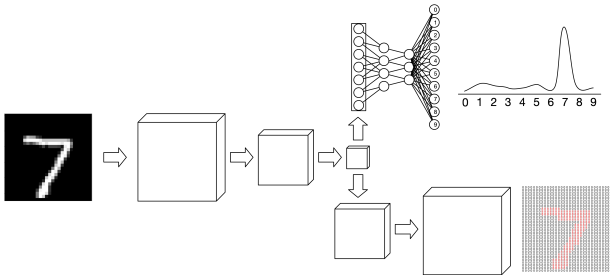
$$f(x_k; \theta_f) = z_k \tag{5.1}$$



(a) Classification Via Segmentation (CvS)



(b) Standard Classification Network



(c) Multi-task Learning Network

Figure 19: This figure illustrates the overall architecture of (a) main network schema for CvS in comparison to (b) standard vanilla classification network, and (c) a standard multi-task learning network.

$$g(z_k; \theta_g) = h_k \tag{5.2}$$

where $f(.)$ parameterized by θ_f represents the backbone function $f : \mathcal{X} \rightarrow \mathcal{H}$ mapping the input image to the latent feature map z_k. $g(.)$ parameterized by θ_g represents the head function $g : \mathcal{H} \rightarrow \mathcal{S}$ mapping the feature encoding z_k to the segmentation map, h_k of size $[H, W, P + 1]$. The extra class (class zero) in $p+1$ represents the background and is discarded. Then we define another function $q(h_k) = y_k$ that averages over the remaining P segmentation maps followed by a softmax function to obtain the class label for the given input. Given the input $x_k \in \mathcal{X}$, function f_c can be decomposed such that

$$f_c(x_k, \theta_c) = (q \circ g \circ f)(x_k) \tag{5.3}$$

where $\theta_c = \{\theta_g, \theta_f\}$. The backbone function $f(.)$ is modeled by a neural network and the head network consists of a stack of convolutional blocks.

Different from standard multi-task learning methods, CvS learns the classification and segmentation tasks simultaneously with only one loss function which we take to be a cross entropy loss.

5.3.2 CvS Segmentation

Motivated by difficluty of procuring segmentation labels, we employ two simple approaches, binarization and label propagation to procure segmented datasets. Although label propagation has been widely used, we employ it in the context of our work by learning a preliminary model from small subset of the dataset ($\sim 1 - 5$ samples per class) with segmentation label and use this model to propagate segmentation labels to the rest of the dataset. This allows us to apply

CvS on the whole dataset of any size by collecting segmentation labels for an extremely small subset.

5.3.2.1 Binarization

For datasets that consists of black and white images such as MNIST, we opted for simplicity and applied a binarization technique with threshold 0.0 to obtain the segmentation maps. In the obtained image, the 0.0 pixel values represent background and the non-zero pixels (pixels with value 1) represent the class of the given image.

5.3.2.2 Segmentation Propagation

For datasets with more complex image data such as CIFAR10/100 the binarization method was no longer applicable. We employed label propagation technique in which we start the algorithm by a very small subset of data being manually segmented and these segmentation labels are propagated to the images without segmentation label. The overall pipeline for segmentation propagation is illustrated in Fig.Figure 20. First, we manually segment M images per class. Then we use the $M \times n_{classes}$ training samples to train the segmentation module of the CvS framework (the functions f and g) which we refer to as Seg-M network. Then we use the trained Seg-M network to predict the segmentation maps for the remaining images.

The binarization and label propagation methods allows us to minimize the tedious work of manual image segmentation.

5.4 Model Architecture

The CvS framework derives the classification label for the given input image via a segmentation module which itself is composed of two main components, backbone and head. The

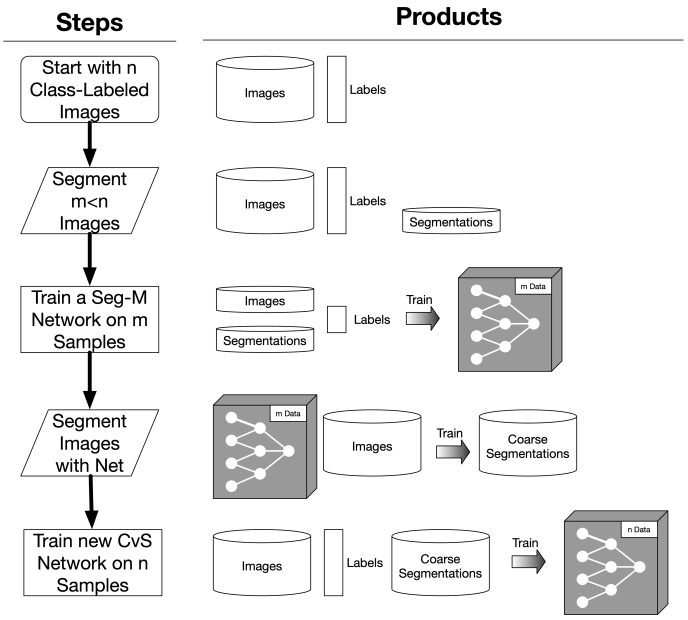


Figure 20: Pipeline for Segmentation Propagation. The pipeline use the segmentation network of CvS trained on few samples with segmentation label, to propagate segmentation labels to the whole dataset.

backbone network takes the input image and learns the latent feature maps which is being utilized by the head network to predict segmentation maps. The predicted segmentation maps are further used to classify the input image into one of the predefined classes. The overall architecture is illustrated in Fig.19(a).

5.4.1 Backbone Architecture

The CvS framework allows various choices of backbone architecture without any constraints. We adopted two of the most commonly used network ResNet-101 and Wide-ResNet for our work.

We employed a custom Wide-ResNet [1] with the depth and width set to 28 and 10 respectively. The network is composed of a convolutional layer followed by a stack of three ResNet block where each consists of two BatchNormalization-ReLU-Conv structure. The first layer in each ResNet block is followed by Dropout. For ResNet-101, we employed a standard architecture used in TorchVision package.

5.4.2 Head Architecture

5.4.2.1 CvS Head

The CvS head is built upon convolutional layer. The architecture of the convolutional head varies slightly depending on the backbone network being used. When Wide-ResNet is used as the backbone network, the head follows the BatchNormalization-ReLU-ConvTranspose structure to output the corresponding segmentation maps. When ResNet101 is used, the DeepLab-Head architecture is employed as the head layer.

5.4.2.2 Baseline Head

To show the effectiveness of our proposed framework, we compared CvS with traditional classifier networks and multi-task learning. These baseline methods has similar same structure to CvS with ResNet101 or Wide-ResNet as their backbone networks.

Linear Head: The linear head is employed in classification networks and follows the BatchNormalization-ReLU-AveragePooling structure followed by linear layer.

---

[1] We borrowed the same network architecture developed by Bumsoo Kim https://github.com/meliketoy and replicated all the experiments for the purpose of our work.

Multi-task Head The multi-task head is employed in multi-task learning method which is composed of two complementary tasks, segmentation and classification. The method consist of two heads outputting the predicted segmentation maps and predicted classification label. When using Wide-ResNet as backbone architecture, we employed a ConvTranspose layer after applying BatchNormalization-ReLU to estimate the corresponding segmentation maps. For the classification head, we applied BatchNormalization-ReLU-AveragePooling followed by a linear layer. For methods using ResNet101 as their backbone network, we employed a stack of three ConvTranspose layer to predict the corresponding segmentation maps where the first two layers are followed by ReLu and BatchNorm. For classification we applied an average pooling layer followed by a linear layer.

5.5 Experiments

5.5.1 Data Collection

In order to evaluate our proposed approach, we employed four datasets, MNIST [46], CIFAR10, CIFAR100 [36] and HRF [66] in the context of image classification.

MNIST: MNIST dataset comprises $60K$ 28×28 hand written digits labeled with 10 classes corresponding to digits 0 to 9. This dataset doesn't contain segmentation labels. Therefore, we employed the binarization technique introduced in previous section to obtain fully segmented data. Given the $[0, 1]$ valued images, zero valued pixels were labeled as background and pixels with value 1 were multiplied by their corresponding class label, i.e. 1×7 representing the class of digit 7. Further, to distinguish the class of digit zero and the background class (zero-valued pixels), we incremented the non-zero pixel values by one i.e. $1 \times 7 + 1$ representing the class

of digit 7. So the class of digit k, $k \in \{0, .., 9\}$ is represented by pixel value of $k + 1$ in its corresponding segmentation map.

CIFAR10/100: The CIFAR-10 dataset consists of $60K$ 32×32 color images with 10 classes including $\{Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, Truck\}$. CIFAR-100 is like CIFAR-10 except it has 100 classes with 600 images per class. These datasets have no segmentation labels. Since the binarization method is not applicable for these datasets, we employed the label propagation approach introduced in the previous section to segment the whole data. To minimize the laborious work of manual segmentation further, we started with CIFAR-10 dataset and manually segmented M images per class where $M \in \{1, 5, 10, 25, 50, 100\}$. Then we trained the Seg-M model using the $M \times 10$ manually segmented images. Then out of the six trained Seg-M networks, we chose Seg-10 and Seg-100 to segment the rest of the images in two different settings for the sake of comparison.

Due to similarity of CIFAR-100 and CIFAR-10, we did not manually segmented any of the images for CIFAR-100. Instead, we employed the Seg-10 and Seg-100 networks from CIFAR-10 to segment the whole CIFAR-100 dataset.

HRF: This dataset consists of 45 High Resolution Fundus images belonging to 45 patients in three classes, healthy, diabetic retinopathy or glaucomatous with 15 images per class. Fundus images represent the posterior part of the eye and are employed mainly for diagnosis of retinal diseases. This dataset also provides the binary gold standard vessel segmentation for each image which eliminated the need for binarization or label propagation methods. A random sample of Fundus photo and its corresponding vessel segmentation from HRF is illustrated in Figure 21.

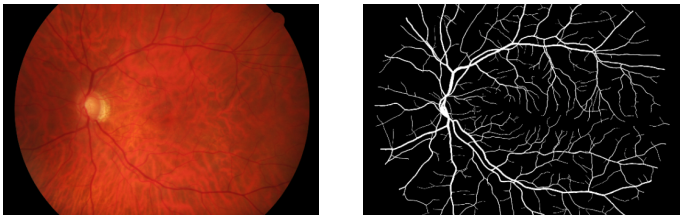


Figure 21: Sample of Fundus photo and its vessel segmentation from HRF.

5.5.2 Experimental Setting

We employ slightly different experimental setting for training CvS based on the choice of backbone network architecture.

Training and Optimization: For CvS with Wide-ResNet architecture as its backbone network, we trained the model from random initialization. Image inputs were kept in their original resolution for MNIST, CIFAR-10 and CIFAR-100, 28×28, 32×32,and 32×32 respectively. For HRF, images were resized to 128×128.

For CvS with ResNet101 architecture as its backbone network, we trained the model using the pretrained network's weights as initialization. Image inputs were resized to 128×128 For MNIST, CIFAR-10, CIFAR-100. For HRF, images were resized to 256×256. The reason for choosing higher image size is the Max-Pooling operation in ResNet-101 and a more fair comparison to Wide-ResNet.

To show the effectiveness of our approach in small data regime, we select small subset of data from each dataset. Hence, the batch size is selected from $8, 16, 32, 128$ depending on the

number of training samples available to the model. For optimizer we use SGD optimizer with momentum parameter set to 0.9, weight decay of 0.0005 and a initial learning rate of 0.1.

Data Augmentation: We explored various combination of data augmentations ranging from resizing to color distortion, shift, rotation, flipping, and adding noise. In this section, we report the data augmentations that achieved the best results for each dataset and the model being used (CvS or baselined methods).

For CvS network, when using MNIST, a random turn, random shift with zoom and gaussian noise are employed for data augmentation. When using CIFAR-10, a random turn, color distortion and random flip are used. When using CIFAR-100, a color distortion, random shift with zoom and random flip are used. When using HRF, random turn and random horizontal flip are used.

For baseline methods (standard classification and multi-task), when using CIFAR-10/100 and MNIST, a random crop and resize with random horizontal flip are employed for data augmentations. When HRF is used, only resize with random horizontal flip are used.

5.5.3 Experimental Results

To show the effectiveness of our proposed method in handling small datasets, we chose different number of training samples ranging from only one sample per class to using the full dataset to train our model. We demonstrate the result of our work on all four datasets in the context of image classification and compare it with previous works.

5.5.3.1 MNIST Performance Analysis

In this section, we demonstrate the result of our work on MNIST dataset. First, we select M random samples per class as our training set where $M \in \{1, 5, 10, 25, 50, 100, 500, 1000, N\}$ and N represents the size of the dataset. We compare our proposed method against a standard classification network and multi-task learning methods proposed in previous sections. Further, we compare our result with previously proposed method LeNet [51]. The result is shown in Table X. The result in Table X, shows that CvS outperforms all the baseline methods where the performance gap is in particular more noticeable the size of the dataset is very small ($M \leqslant 100$). CvS achieves comparable results to multi-task learning when the number of training samples is larger than $50k$. We can also see that using ResNet-101 as the backbone network achieves superior result over the Wide-ResNet when the model has access to only $M = 1$ example per class and achieves comparable results for dataset size of larger than 50 ($M \geqslant 5$).

The overall results indicates the effectiveness of our proposed approach for classification with only a handful of training samples. Our proposed CvS framework can achieve high performance without the need for large amount of data for pre-training or significant amount of computational resources.

5.5.3.2 CIFAR-10 Performance Analysis

Similar to the MNIST experimental setting, we selected $M \in \{1, 5, 10, 25, 50, 100, 500, 1000, N\}$ random samples per class as the training data where N represents the size of the dataset. Table Table XI compares the results of our work with the standard classification network and two

TABLE X: Classification accuracy on MNIST Test set given different numbers of samples per class and methodology.

Methods	Number of samples per class								
	1	5	10	25	50	100	500	1k	Full
Backbone:ResNet101									
Classification	58.21	50.38	68.09	82.78	91.09	94.06	98.62	98.68	99.47
Multi-task	48.82	79.43	83.26	93.94	95.19	96.65	**99.36**	99.39	**99.75**
CvS	**71**	**87.67**	**92.7**	**95.59**	**97.8**	**97.8**	99.11	99.17	99.62
Backbone:W-ResNet									
Classification	20.16	77.05	78.78	84.88	89.31	95.45	98.49	98.83	99.16
Multi-task	15.25	30.87	37.17	84.6	96.28	97.87	**99.18**	**99.51**	99.45
CvS	**54**	**88**	**90.9**	**95.68**	**97.41**	**97.95**	99.08	99.25	**99.51**
Other Architecture									
LeNet	47.7	-	72	-	-	82	-	-	98.5

of the previously proposed approaches Big Transfer [63] and Deep Metric Transfer [64]. The results for Big Transfer and Deep Metric Transfer are reported directly from their papers.

Table XI shows the result for CvS where $M \in \{1, 5, 10, 25, 50, 100\}$ samples per class were manually segmented. Then we employed label propagation technique that was introduced in previous section to segment the remaining images. First, we trained the CvS model using the images and their corresponding manually segmentation labels. We then selected two of the

trained CvS models for each of the backbone network choices and employed their corresponding segmentation network, Seg-M to propagate the segmentation label to the unlabeled images. We chose Seg-10 and Seg-100 that are trained with 10 and 100 samples per class respectively and are depicted by $*$ and $**$ in Table Table XI. The CvS(Seg-10) and CvS(Seg-100) in Table Table XI indicates the performance of CvS framework when Seg-10 and Seg-100 were employed to obtain the segmentation labels.

As Table Table XI suggests, CvS outperforms traditional classification networks and Deep Metric Transfer [64] significantly. The performance gap is particularly noticeable in low data regime. This result suggests that CvS is much more powerful in tackling the overfitting problem when dealing with small datasets as opposed to traditional classifiers. We can also see that CvS achieves its best performance when using ResNet-101 as the backbone network.

As it was expected, CvS models that have access to manually segmented images performs slightly better than those with predicted segmented data. Comparing the results from CvS(Seg-10) and CvS(Seg-100) shows that increasing the number of manually segmented images does not benefit the model significantly. This result helps with reducing the cost of laborious manual segmentation and makes our proposed approach a cost-effective method that is able to achieve high performance with access to only a handful of labeled data.

Although CvS doesn't perform as well as Big Transfer, there are few limitations with this method that limits its application. First, Big Transfer exploits the weights of pre-trained network on the assumption of availability of large amount of data for pre-training. Second, transfer learning may not work to it full potential when its transferred to a new task where the

TABLE XI: Classification accuracy on CIFAR10 Test set given different numbers of samples per class and methodology.

Methods	Number of samples per class								
	1	5	10	25	50	100	500	1k	Full
Backbone:ResNet101									
Classification	18.85	20.67	26.01	27.12	37.37	42.16	69.75	76.4	93.55
CvS	**39.31**	**67.24**	**73.94***	**80.95**	**86.43**	**90.1***	-	-	-
CvS(Seg-10)	-	-	-	78.69	84.89	88.51	93.26	94.66	96.42
CvS(Seg-100)	-	-	-	-	-	-	**93.79**	**95.37**	**97.13**
Backbone:W-ResNet									
Classification	16.66	25.03	26.11	35.47	42.34	54.2	79.29	85.49	93.71
CvS	19.35	33.07	38.56**	51.4	59.69	68.5**	-	-	-
CvS(Seg-10)	-	-	-	45.45	54.09	62.12	78.93	84.81	93.29
CvS(Seg-100)	-	-	-	-	-	-	80.76	85.81	93.18
Other Architectures									
Big Transfer	**67**	**94**	**97**	-	-	**98**	-	-	**99.4**
DeepMetric-Transfer	-	56.3	63.5	-	74.8	79.4	84.6	87.9	-

nature of data is very different from the data used for pre-training. Additionally, training large networks on large datasets that their weights can be later used, requires powerful computational resources that may not be accessible to everyone.

5.5.3.3 CIFAR-100 Performance Analysis

We chose a similar experimental set up for CIFAR-100. We selected $M \in \{1, 5, 10, 25, 50, 100, N\}$ random samples per class where N represents the size of dataset. Since we did not segment any of the CIFAR-100 images manually, we employed the Seg-10 and Seg-100 networks from the CvS model trained on CIFAR-10 (depicted by $*$ in Table Table XI) to obtain fully segmented data for CIFAR-100 images. Table Table XII compares the results of our work with the standard classification network and Big Transfer [63]. The observation from Table Table XII supports the results in Table Table XI. The result shows the superiority of CvS over standard classifier networks. We can also see that the model doesn't benefit from increasing manually segmented images. The overall result indicates that the CvS model can achieve much higher performance than vanilla classification network with negligible cost in manual segmentation.

5.5.3.4 HRF Performance Analysis

In this section, we evaluate our work on a medical dataset for a real-world application of ophthalmic disease classification and compare it with standard classification networks, multi-task learning, and previous work from [67]. The result is shown in Table Table XIII.

After 5-fold cross validation on our 45 images (15 images per class), we get a final accuracy of 82% which outperforms all the baselines. Each fold of the cross validation used 36 images in the training set and was trained on the left-out 9 images. Cross Validation folds were chosen randomly and there was no validation/development set, which meant hyperparameters and network stopping point were not selected for. The hyperparameters that worked best for the CIFAR-10 experiments were used.

TABLE XII: Classification accuracy on CIFAR100 Test set given different numbers of samples per class and methodology. All segmentation labels were propagated from CIFAR10 trained networks, following the same nomenclature as the section above.

Methods	Number of samples per class						
	1	5	10	25	50	100	Full
Backbone:ResNet101							
Classification	2.97	7.6	11.25	24.9	38.2	55.48	78.24
CvS (Seg-10)	**21.78**	**46.93**	**56.4**	**65.21**	**70.24**	**75.14**	**83.9**
CvS (Seg-100)	21.49	45.73	52.8	64.07	70.23	74.65	83.64
Backbone:Wide-ResNet							
Classification	3.89	7.12	9.38	31.45	33.75	55.85	78.01
CvS (Seg-10)	10.23	19.35	24.54	33.51	41.89	53.02	75.76
CvS (Seg-100)	10.79	20.03	25.11	33.45	40.75	51.46	72
Other Architectures							
Big Transfer (SoTA)	**40**	**78**	**84**	**87**	-	**91**	**93.5**

As Table Table XIII suggests, CvS outperforms all the baselines. Despite the high performance reported by [67], this method does not use full Fundus but it crops the image around the optic disc. It also uses a combination of 5 datasets as training data and evaluate it on HRF which limits the method for applications with extremely small datasets.

TABLE XIII: Classification accuracy on HRF for disease detection.

Methods	Accuracy(%)
Backbone:ResNet101	
Classification	66.67
Multi-task	70.23
CvS (ours)	**82.22**
Other Architectures	
Xception [67]	80.00

5.5.4 Label Propagation Analysis

In this section, we analyze the quality of predicted segmentation maps performed by label propagation technique. We visualize the result of Seg-10 and Seg-100 networks of CvS framework on a sample image selected randomly from each class of CIFAR-10 and CIFAR-100. For CIFAR-10, we further compare the result against manually segmented images. The result for CIFAR-10 and CIFAR-100 are shown in Figure 22 and Figure 23 respectively. In Figure 22, from left to right, we have an original image randomly selected from each class and their corresponding segmentation maps: (a) represents manual segmentation, (b) and (c) represent predicted segmentation performed by Seg-10 and seg-100 networks of CvS respectively. In Figure 23, both Seg-10 and Seg-100 are borrowed from CvS that was trained with CIFAR-10.

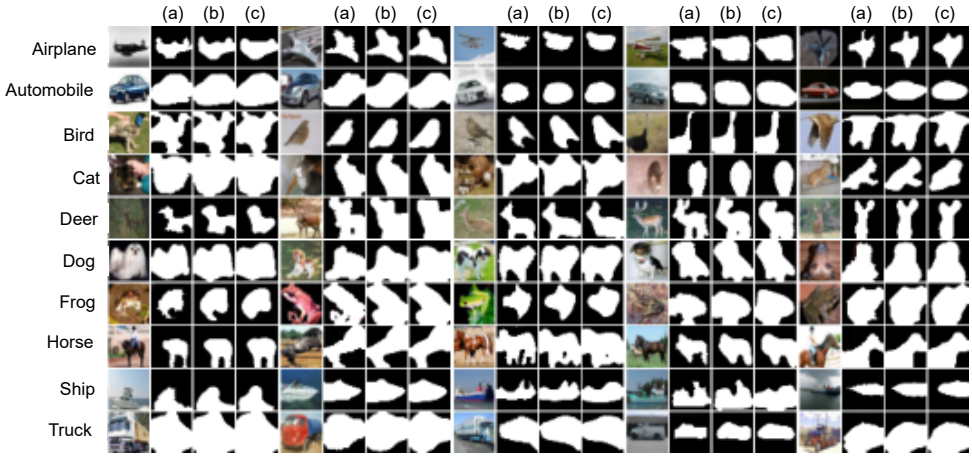


Figure 22: Illustration of predicted segmentation maps performed by label propagation technique for CIFAR-10.

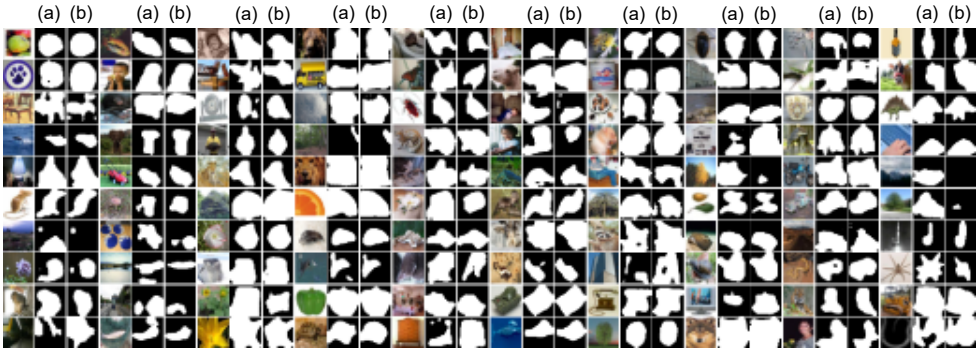


Figure 23: Illustration of predicted segmentation maps performed by (a) Seg-10 and (b) Seg-100 networks of CvS model for CIFAR-100.

As can be seen, the proposed model shows a reasonable accuracy in predicting the segmentation maps for both CIFAR-10 and CIFAR-100. The results suggests that the label propagation technique, can achieve a fully segmented dataset with access to only handful of data points being manually segmented. Therefore, the cost of laborious manual segmentation reduces and collecting segmented datasets become more accessible.

5.6 Conclusion

In this chapter we study the problem of learning from small datasets for image classification tasks. Deep neural networks performs to their full potential in modeling and problem solving when trained with large amount of data. However, labels are expensive, difficult or infeasible to obtain. On the other hand, classification network tend to overfit when trained on only a small amount data.

Current approaches mainly address the small data problem by formulating the problem in transfer learning or multi-task learning settings. Transfer learning methods ususally performs to their full potential when the nature of data in the downstream task is similar to the one used in the pretrained models. Multi-task learning face the problem of difficulty of obtaining segmentation label and loss balancing in different heads in the model often leading to overfitting. In this chapter, we propose a novel framework to the above problem that harnesses the power of segmentation to learn from small datasets enabling classification on extremely small datasets ($\tilde{1}$-5 samples per class). As opposed to standard multi-task learning, our proposed framework, called CvS, is a single headed model eliminating the problem of loss balancing in different head. We also propose to employ binarization and label propagation methods to obtain segmentation

labels in a cost-effective way. The label propgation allows us to obtain a fully segmented data with segmenting only a small subset of data ($\tilde{1}$-5 samples per class). We studied the component of our method and showed its effectiveness on a broad range of image classification tasks. Our experiments showed considerable improvement over vanilla classification network and multi-task learning.

APPENDICES

Copyright

In all INSTICC publications, the copyright to the contribution is transferred from the Author to the Publisher, i.e. Science and Technology Publications, (SCITEPRESS) Lda. The copyright transfer covers the rights to reproduce and distribute the contribution, including reprints, translations, photographic reproductions, microform, electronic form (offline, online), or any other reproductions of similar nature, including book publication.

The Author retains the rights to publish the contribution in his/her own website or in his/her employer's website, as long as it is clearly stated in which publication or event it was originally published in and a link to the original publication or event is made.

The Author warrants that his/her contribution is original, except for such excerpts from copyrighted works as may be included with the permission of the copyright holder and author thereof, that it contains no libelous statements and does not infringe on any copyright, trademark, patent, statutory right, or propriety right of others. The Author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors.

In return for these rights, the publisher agrees to have the identified contribution published, at its own cost and expense, in the publication or conference proceedings the paper is submitted to.

Please note: This is a shortened version of the copyright agreement which authors will sign when publishing at an INSTICC event or publication. This agreement can vary and it is important to read the specific copyright agreement for each event or publication thoroughly.

COMMUNITY

- Research Community
- Membership
- News
- Distinguished Researchers
- Keynotes & Interviews

PUBLICATIONS

- Reasons to Publish
- Digital Library
- Author Resources
  - Paper Submission
  - Plagiarism Policy
  - Copyright
  - Indexation
  - Paper Presentation

EVENTS

- Event Structure
- Submission Deadlines
- Attend an Event
  - As a Speaker
  - As a Non-Speaker
  - As an Exhibitor
- Propose an Event
- Information System

PARTNERSHIPS

- Partnership Opportunities
- Academic Partners
- Industrial Partners
- Institutional Partners
- Publication Partners

ABOUT US

- Mission and Activities
- Contacts
- Privacy Policy

Deep Multi-Task Learning for Interpretable Glaucoma Detection

Conference Proceedings:
2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)

Author: Nooshin Mojab

Publisher: IEEE

Date: July 2019

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW

CITED LITERATURE

1. Mojab, N., Noroozi, V., Aleem, A., Nallabothula, M. P., Baker, J., Azar, D. T., Rosenblatt, M., Chan, R. V. P., Yi, D., Yu, P. S., and Hallak, J. A.: I-oda, real-world multi-modal longitudinal data for ophthalmic applications. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021, Volume 5: HEALTHINF, Online Streaming, February 11-13, 2021* , eds. C. Pesquita, A. L. N. Fred, and H. Gamboa, pages 566–574. SCITEPRESS, 2021.

2. Mojab, N., Noroozi, V., Yu, P., and Hallak, J.: Deep multi-task learning for interpretable glaucoma detection. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)* , pages 167–174. IEEE, 2019.

3. Mojab, N., Noroozi, V., Yi, D., Nallabothula, M. P., Aleem, A., Philip, S. Y., and Hallak, J. A.: Real-world multi-domain data applications for generalizations to clinical settings. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* , pages 677–684. IEEE, 2020.

4. Mojab, N., Yu, P., Hallak, J., and Yi, D.: Cvs: Classification via segmentation for small datasets. In *Proceedings of the 32th British Machine Vision Conference (BMVC 2021)* , 2021, submitted.

5. Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B. S., Waldstein, S. M., and Bogunović, H.: Artificial intelligence in retina. *Progress in retinal and eye research* , 67:1–29, 2018.

6. Lu, W., Tong, Y., Yu, Y., Xing, Y., Chen, C., and Shen, Y.: Applications of artificial intelligence in ophthalmology: general overview. *Journal of ophthalmology* , 2018, 2018.

7. Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., and Wong, T. Y.: Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology* , 103(2):167–175, 2019.

8. Grewal, P. S., Oloumi, F., Rubin, U., and Tennant, M. T.: Deep learning in ophthalmology: a review. *Canadian Journal of Ophthalmology* , 53(4):309–313, 2018.

9. Burlina, P. M., Joshi, N., Pekala, M., Pacheco, K. D., Freund, D. E., and Bressler, N. M.: Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA ophthalmology* , 135(11):1170–1176, 2017.

10. Burlina, P. M., Joshi, N., Pacheco, K. D., Freund, D. E., Kong, J., and Bressler, N. M.: Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. *JAMA ophthalmology* , 136(12):1359–1366, 2018.

11. Varadarajan, A. V., Poplin, R., Blumer, K., Angermueller, C., Ledsam, J., Chopra, R., Keane, P. A., Corrado, G. S., Peng, L., and Webster, D. R.: Deep learning for predicting refractive error from retinal fundus images. *Investigative ophthalmology & visual science* , 59(7):2861–2868, 2018.

12. Gargeya, R. and Leng, T.: Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* , 124(7):962–969, 2017.

13. Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama* , 316(22):2402–2410, 2016.

14. Medeiros, F. A., Jammal, A. A., and Thompson, A. C.: From machine to machine: an oct-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs. *Ophthalmology* , 126(4):513–521, 2019.

15. Thompson, A. C., Jammal, A. A., and Medeiros, F. A.: A deep learning algorithm to quantify neuroretinal rim loss from optic disc photographs. *American journal of ophthalmology* , 201:9–18, 2019.

16. Fu, H., Cheng, J., Xu, Y., Zhang, C., Wong, D. W. K., Liu, J., and Cao, X.: Disc-aware ensemble network for glaucoma screening from fundus image. *IEEE transactions on medical imaging* , 37(11):2493–2501, 2018.

17. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* , pages 248–255. Ieee, 2009.

18. Fumero, F., Alayón, S., Sanchez, J. L., Sigut, J., and Gonzalez-Hernandez, M.: Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)* , pages 1–6. IEEE, 2011.

19. Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., and Lakshminarayanan, V.: Retinal fundus images for glaucoma analysis: the riga dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications* , volume 10579, page 105790B. International Society for Optics and Photonics, 2018.

20. Sivaswamy, J., Krishnadas, S., Joshi, G. D., Jain, M., and Tabish, A. U. S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)* , pages 53–56. IEEE, 2014.

21. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology* , 33(3):231–234, 2014.

22. Mancino, R., Martucci, A., Cesareo, M., Giannini, C., Corasaniti, M. T., Bagetta, G., and Nucci, C.: Glaucoma and alzheimer disease: one age-related neurodegenerative disease of the brain. *Current neuropharmacology* , 16(7):971–977, 2018.

23. Davis, B. M., Crawley, L., Pahlitzsch, M., Javaid, F., and Cordeiro, M. F.: Glaucoma: the retina and beyond. *Acta neuropathologica* , 132(6):807–826, 2016.

24. Tham, Y.-C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., and Cheng, C.-Y.: Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* , 121(11):2081–2090, 2014.

25. Khan, F., Khan, S. A., Yasin, U. U., ul Haq, I., and Qamar, U.: Detection of glaucoma using retinal fundus images. In *The 6th 2013 Biomedical Engineering International Conference* , pages 1–5. IEEE, 2013.

26. Sevastopolsky, A.: Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognition and Image Analysis* , 27(3):618–624, 2017.

27. Zilly, J. G., Buhmann, J. M., and Mahapatra, D.: Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images. In *International Workshop on Machine Learning in Medical Imaging* , pages 136–143. Springer, 2015.

28. Chen, X., Xu, Y., Wong, D. W. K., Wong, T. Y., and Liu, J.: Glaucoma detection based on deep convolutional neural network. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* , pages 715–718. IEEE, 2015.

29. Abbas, Q.: Glaucoma-deep: Detection of glaucoma eye disease on retinal fundus images using deep learning. *International Journal of Advanced Computer Science and Applications* , 8(6):41–45, 2017.

30. Raghavendra, U., Fujita, H., Bhandary, S. V., Gudigar, A., Tan, J. H., and Acharya, U. R.: Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Information Sciences* , 441:41–49, 2018.

31. Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* , pages 234–241. Springer, 2015.

32. Zilly, J., Buhmann, J. M., and Mahapatra, D.: Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics* , 55:28–41, 2017.

33. Gómez-Valverde, J. J., Antón, A., Fatti, G., Liefers, B., Herranz, A., Santos, A., Sánchez, C. I., and Ledesma-Carbayo, M. J.: Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomedical optics express* , 10(2):892–913, 2019.

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

35. Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* , 2015.

36. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. 2009.

37. Bengio, Y.: Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning* , pages 17–36, 2012.

38. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* , 2020.

39. Doersch, C., Gupta, A., and Efros, A. A.: Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision* , pages 1422–1430, 2015.

40. Noroozi, M. and Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision* , pages 69–84. Springer, 2016.

41. Gidaris, S., Singh, P., and Komodakis, N.: Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* , 2018.

42. Zhang, R., Isola, P., and Efros, A. A.: Colorful image colorization. In *European conference on computer vision* , pages 649–666. Springer, 2016.

43. Kolesnikov, A., Zhai, X., and Beyer, L.: Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* , pages 1920–1929, 2019.

44. You, Y., Gitman, I., and Ginsburg, B.: Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888* , 2017.

45. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L.: Microsoft coco: Common objects in context. In *European conference on computer vision* , pages 740–755. Springer, 2014.

46. LeCun, Y., Cortes, C., and Burges, C.: Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* , 2, 2010.

47. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q.: A comprehensive survey on transfer learning. *Proceedings of the IEEE* , 109(1):43–76, 2020.

48. Zhang, Y. and Yang, Q.: An overview of multi-task learning. *National Science Review* , 5(1):30–43, 2018.

49. Zhang, Y. and Yang, Q.: A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* , 2021.

50. Badrinarayanan, V., Kendall, A., and Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* , 39(12):2481–2495, 2017.

51. LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D.: Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems* , pages 396–404, 1990.

52. Krizhevsky, A., Sutskever, I., and Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* , pages 1097–1105, 2012.

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

54. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 1–9, 2015.

55. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z.: Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 2818–2826, 2016.

56. He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 770–778, 2016.

57. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q.: Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* , pages 4700–4708, 2017.

58. Pham, H., Dai, Z., Xie, Q., Luong, M.-T., and Le, Q. V.: Meta pseudo labels. *arXiv preprint arXiv:2003.10580* , 2020.

59. Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412* , 2020.

60. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* , 2020.

61. Byerly, A., Kalganova, T., and Dear, I.: A branching and merging convolutional network with homogeneous filter capsules. *arXiv preprint arXiv:2001.09136* , 2020.

62. Mazzia, V., Salvetti, F., and Chiaberge, M.: Efficient-capsnet: Capsule network with self-attention routing. *arXiv preprint arXiv:2101.12491* , 2021.

63. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N.: Big transfer (bit): General visual representation learning. *arXiv preprint arXiv:1912.11370* , 2019.

64. Liu, B., Wu, Z., Hu, H., and Lin, S.: Deep metric transfer for label propagation with limited annotated data. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* , pages 0–0, 2019.

65. Zhang, W., Li, R., Zeng, T., Sun, Q., Kumar, S., Ye, J., and Ji, S.: Deep model based transfer and multi-task learning for biological image analysis. *IEEE transactions on Big Data* , 6(2):322–333, 2016.

66. Budai, A., Bock, R., Maier, A., Hornegger, J., and Michelson, G.: Robust vessel segmentation in fundus images. *International journal of biomedical imaging* , 2013.

67. Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., and Navea, A.: Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online* , 18(1):1–19, 2019.

VITA

NAME: NOOSHIN MOJAB

EDUCATION: Ph.D., Computer Science, University of Illinois at Chicago, Chicago, Illinois, 2021.

B.Sc., Computer Science, Sharif University of Technology, Tehran, Iran, 2010.

ACADEMIC EXPERIENCE: Research Assistant, Artificial Intelligence in Ophthalmology Lab (Ai-O), Department of Ophthalmology and Visual Sciences, University of Illinois at Chicago, 2018 - 2021.

Teaching Assistant, Department of Computer Science, University of Illinois at Chicago:

- Mathematical Foundation of Computing, Spring 2015 and Summer 2017.
- Program Design I, Fall 2015 and Fall 2016.
- Database System, Spring 2016
- Software Engineering, Spring 2017, Fall 2017, Spring 2018.