Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

4-18-2023 12:00 PM

Open-Set Source-Free Domain Adaptation in Fundus Images Analysis

Masoud Pourreza, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science

© Masoud Pourreza 2023

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Other Computer Engineering Commons

Recommended Citation
Pourreza, Masoud, "Open-Set Source-Free Domain Adaptation in Fundus Images Analysis" (2023). *Electronic Thesis and Dissertation Repository*. 9241.
https://ir.lib.uwo.ca/etd/9241

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

Unsupervised domain adaptation (UDA) is crucial in medical image analysis where only the source domain data is labeled. There is a lot of emphasis on the closed-set paradigm in UDA, where the label space is assumed to be the same in all domains. However, medical imaging often has an open-world scenario where the source domain has a limited number of disease categories and the target domain has unknown distinct classes. Also, maintaining the privacy of patients is a crucial aspect of medical research and practice. In this work, we shed light on the Open-Set Domain Adaptation (OSDA) setting on fundus image analysis while preserving the privacy concern. In particular, we step towards a source-free open-set domain adaptation where, without source data, the source model is utilized to facilitate adaptation to open-set unlabeled data by delving into channel-wise and local features for fundus disease recognition. In particular, considering the nature of the fundus images, we present a novel objective way in the adaptation phase to utilize spatial and channel-wise information to select the best source model for a target domain, even by considering the small inter-class variation between samples. Our approach has achieved state-of-the-art performance compared to other methods.

Keywords: Domain Adaptation, Fundus Images Analysis, Open-Set Domain Adaptation, Source-Free Domain Adaptation

Summary for Lay Audience

Medical researchers often want to use data from one group of patients (source domain) to understand diseases in another group of patients (target domain). However, this can be difficult when the data from the target domain doesn't have labels that tell us what the diseases are. In the past, researchers have tried to use labels from the source domain to understand the target domain, but this only works if the diseases in both groups are the same. But sometimes the diseases in the target domain are different from those in the source domain. To solve this problem, researchers have developed a method called open-set domain adaptation. This method allows them to use data from the source domain to understand the target domain without actually looking at the data from the source domain. This is important because it helps protect the privacy of patients. In this study, we apply source-free domain adaptation to understand various types of eye diseases.

Contents

Abstract i

Summary for Lay Audience ii

List of Figures vi

List of Tables viii

List of Abbreviations ix

1 Introduction 1
- 1.1 Motivation 1
- 1.2 Clinical Problem Statement 2
  - 1.2.1 Retinal Fundus Images 2
    - Eye's Anatomy 2
    - Retinal Fundus Photography 2
    - Ocular Diseases 3
  - 1.2.2 Fundus Images Characteristics 4
  - 1.2.3 Clinical Challenges 5
- 1.3 Our Solution 5
  - 1.3.1 Unsupervised Domain Adaptation 6
  - 1.3.2 Open-Set Recognition 6
  - 1.3.3 Open-Set Domain Adaptation 7
- 1.4 Contributions 8
- 1.5 Thesis Outline 8

2 Background and Literature Review 10
- 2.1 Domain Adaptation 10
  - 2.1.1 Domain Adversarial Training of Neural Network 12
  - 2.1.2 Pseudo-labeling in Domain Adaptation 13
  - 2.1.3 Universal Domain Adaptation 13
- 2.2 Source-Free Domain Adaptation 14
  - 2.2.1 Entropy Minimization and Self-Supervised Learning 14
  - 2.2.2 Generative Adversarial Network domain Domain Adaptation 15
    - Universal Source-Free Domain Adaptation 16
- 2.3 Open-Set Domain Adaptation 16

2.3.1 Instance-Level Approaches . . . 17
Backpropagation-based . . . 17
Multiple Classifiers . . . 18
Progressive Separation . . . 18
2.3.2 Hybrid Approaches . . . 18
Self-Supervision Using Roatation . . . 19
Clustering-Based . . . 19
Source-Free Open-set Domain Adaptation . . . 19
2.4 Fundus Image Analysis . . . 20
2.4.1 Deep Learning Ensemble Approach . . . 20
2.4.2 Transformers for Fundus Image Classification . . . 21
2.4.3 Graph Neural Network for Fundus Image Classification . . . 21
2.4.4 Domain Adaptation for Fundus Images . . . 22
2.5 Attention . . . 22
2.5.1 Soft attention . . . 22
Spatial attention . . . 23
Channel attention . . . 23
Self attention . . . 23
2.5.2 Hard attention . . . 24
2.6 Summary . . . 24

3 Methodology **25**
3.1 Problem Formulation . . . 25
3.2 Method Overview . . . 25
3.3 Source model training . . . 26
3.4 Target model adaptation . . . 28
3.4.1 Inherit . . . 28
3.4.2 Tune . . . 28
3.4.3 Spatial and Channel-wise Adaptation (SCA) . . . 29
Integrated Spatial Attention and Channel Attention Module (ISCA) . . . 29
ResNet . . . 29
ResNeXt . . . 30
ResNeSt . . . 31
Channel Attention Module . . . 31
Spatial Attention Module . . . 32
Spatial and Channel-wise objective . . . 33
3.4.4 Summary . . . 34

4 Experimental Analysis **35**
4.1 Experimental Settings . . . 35
4.1.1 Implementation Details . . . 35
4.1.2 Datasets . . . 35
4.1.3 Metrics . . . 37
4.2 Results . . . 37
4.2.1 Comparison Methods . . . 37

4.2.2 Quantitative Results 38
4.2.3 Ablation Study 39
Negative Sample Generation 39
Effect of ISCA Module 40
4.3 Model Interpretation 40
4.3.1 Saliency Map 41
Guided Backpropagation. 41
Gradient-weighted Class Activation Map (Grad-CAM). 42
4.4 Summary 42

5 Conclusion and Future Directions **44**
5.1 Conclusion 44
5.2 Future Directions 45

Bibliography **46**

Curriculum Vitae **51**

List of Figures

1.1 Different channels of a fundus image. The Green channel has more invaluable information. 4

2.1 Domain adaptation in the presence of domain shift. 11

2.2 Domain adaptation settings based on source and target label sets (Shared labels are indicated by colored rectangles). The image is redrawn from [49]. Copyright © 2019, IEEE. 17

3.1 Overview of our method. 26

3.2 We apply feature-splicing by eliminating the class-discriminative characteristics, which involves replacing the activations in the top-(1/3) percentile (d1). The image is redrawn from [24]. Copyright © 2020, IEEE. 27

3.3 (a) ResNeSt represents the height, width, and number of channels of the feature map with h, w, and c. In addition, each cardinal group is divided into r splits.(b) The ResNeSt architecture incorporates a split attention module. Image reprinted from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022. 30

3.4 An overview of our integrated channel attention and spatial attention module. (The image is reprinted from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022.) 31

3.5 (a) Our Spatial Attention (SA). (b) Our Channel Attention (CA). (This figure reproduced from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022.) 32

4.1 ODIR dataset samples and different classes [2]. 36

4.2 RFMiD dataset samples and classes [31]. 37

4.3 TAOP dataset samples and classes [31]. 37

4.4 Saliency map results of fundus images. From left to right, the normal fundus image, the fundus image with red abnormal hemorrhage, the Grad-CAM, and the Guided backpropagation output. The results obtained from both GradCAM and guided backpropagation demonstrate the excellent ability of the model to locate the place of abnormalities or diseases accurately. 41

4.5 Comparing the performance of Resnet and ResNeSt plus attention module on detecting local features. from left to right, the normal fundus image, the fundus image with red abnormal hemorrhage, grad-cam output for Resnet and grad-cam output for the ResNeSt with attention module. ResNeSt, equipped with the proposed module, outperforms the ResNet network in terms of localization of features on fundus images. 42

List of Tables

4.1 Result comparison (%) of state-of-the-art OSDA methods 39

4.2 Adaptation performance (%) of different negative sample generation strategies. OS* and OS represent the mean accuracy over common classes and the mean accuracy over all classes, respectively. UNK is a measure of accuracy in recognizing unknown classes, and HM is a harmonic mean accuracy.The results indicate that feature splicing is more effective than interpolation in generating negative samples. 40

4.3 Adaptation performance (%) of different backbones and the effect of proposed ISCA module. 40

List of Abbreviations

ARMD	Age-Related Macular Degeneration
BRVO	Branch Retinal Vein Occlusion
CRVO	Central Retinal Vein Occlusion
CAM	Class Activation Map
CRCA	Collaborative Regional Clustering and Alignment
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DR	Diabetic Retinopathy
DA	Domain Adaptation
DANN	Domain Adversarial Neural Network
DCC	Domain Consensus Clustering
DSBN	Domain-Specific Batch Normalization
GAN	Generative adversarial Network
OSDA	Open-Set Domain Adaptation
OSBP	Open-Set Domain Adaptation by BackPropagation
OSR	Open-Set Recognition
PM	Pathological Myopia
RVO	Retinal Vein Occlusion
ROS	Rotation-based Open-Set
SSL	Self-Supervised Learning
STA	Separate To Adapt
SFDA	Source-Free Domain Adaptation
UAN	Universal Adaptation Network
UDA	Unsupervised Domain Adaptation
VIT	Vision Transformer

Chapter 1

Introduction

1.1 Motivation

According to the latest World Health Organization (WHO) report on vision, there are 2.2 billion visually impaired people worldwide, and at least 1 billion cases could have been prevented or can still be addressed [4]. Moreover, according to Fighting Blindness Canada (FBC), approximately 8 million Canadians are at risk of losing their vision because of eye diseases [1]. Given the scale of the problem, it is clear that developing accurate and efficient methods for analyzing medical images, such as fundus images, is of great importance. However, one of the major challenges in medical image analysis is the limited availability of annotated datasets, which is often due to concerns about patient privacy.

In such cases, Source-Free Domain Adaptation (SFDA) can be a good choice for fundus image analysis. SFDA involves adapting a model that has been trained on one domain (e.g. a dataset of fundus images) to a new domain (e.g. a different dataset of fundus images) without using any labeled data from the target domain. This is particularly useful in situations where it is difficult or impossible to obtain a large annotated dataset for the target domain, as is often the case in medical image analysis. Another advantage of SFDA is that it allows for the sharing of pre-trained models between different domains. This can be particularly useful in the medical field, where it may be possible to share a pre-trained model across different hospitals or clinics, even if it is not possible to share the underlying data. One variant of domain adaptation is open-set domain adaptation, which involves adapting a model to a target domain where the distribution of the classes in the target domain is not known in advance. This is a common scenario in medical image analysis, where the types of diseases or abnormalities that may be present in the target domain may not be known beforehand.

In summary, the motivation for studying open-set source-free domain adaptation for fundus image analysis is twofold. First, it allows for the adaptation of a model to a new domain without access to labeled data from the target domain, which is often a challenge in medical image analysis due to concerns about patient privacy. Second, it allows for the sharing of pre-trained models between different domains, which can be useful in the medical field where it may not always be possible to share the underlying data. These reasons motivate us to choose this topic as an important and timely research direction.

1.2 Clinical Problem Statement

1.2.1 Retinal Fundus Images

In the human body, vision is arguably the most important sense. The majority of information we receive about the world comes from our eyes, so much so that a significant portion of the brain is completely dedicated to visual processing. Due to the way the eye processes light and converts it into information, it is often compared to a camera. Both have lenses to focus the incoming light. Cameras create images using film, whereas the eye produces images using a layer of specialized cells called the retina. This is where the similarity ends. The eye's ability to focus on a wide range of objects of different sizes, luminosity and contrast at high speed is more potent than current cameras.

Eye's Anatomy

An image is formed by filtering and directing light through the cornea. The pupil adjusts in response to changes in light intensity by means of the iris muscles. In order to focus the light onto the retina, the lens stretches or compresses. Fundus is the inner surface opposite lens, completing the eye's interior structure [15].

The back of the eye hosts the retina, a sensory tissue composed of multiple layers. As light beams pass through the retina, they are converted to electrical signals, which travel along the optic nerve and, finally, to images in the brain. Rods and cones are the two types of photoreceptors in the retina. Even in low light conditions, rod cells can detect changes in contrast and detect motion, but not color. Their main role is to facilitate scotopic vision (night vision). On the other hand, cones are precise cells that detect color. They are primarily found in the macula, which provides photopic vision (day vision) [35].

Retinal Fundus Photography

The retina is the eye's light-sensitive layer covering the eyeball's inner surface. It is responsible for translating the image projected by the lens on its surface into an electrical signal which is then transferred to the brain via nerve fibers.

If we look at the retina's surface through the pupil, the optic disc can be viewed as a bright oval where the veins and arteries extend from its center. It is also possible to see a darker part of the retina, called the macula. This area of the retina is responsible for central vision. There is a dense concentration of cones in the fovea at the center of the macula, but there are no rods there. Seeing through the fovea is made possible by the dense concentration of cones, which is why we move our eyes constantly to bring the target to the center. The foveal avascular zone (FAZ) is the region around the fovea where no vessels cover the retina. The main anatomical parts of the retina that can be observed in this modality are the macula, the fovea, the optic disc, the peripapillary area, and the retinal vasculature. Light passes through a series of lenses and reaches the retina through the pupil. The light reflected from the retina then passes through the pupil again and is collected and guided via lenses and mirrors to form the image of the retina. These cameras are described by their angle of view, usually ranging between 20° to 140°. A camera with a broader angle of view captures a wider image of the retina. On the other hand, by narrowing the angle of view, the resulting image becomes more magnified. To enhance the clarity of photographs, ophthalmologists dilate the pupil before retinal imaging.

Fundus imaging is used to diagnose and monitor various retinal diseases or abnormalities, including diabetic retinopathy, glaucoma, and eye cancer. The availability of fundus photography and the fact that it is noninvasive make it an ideal method of taking multiple images of the retina and assessing any changes that may occur over time [15].

Ocular Diseases

In brief, the followings are a few examples of diseases/abnormalities and their significant visual characteristics.

Diabetic Retinopathy (DR): is one of the most common complications of diabetes, particularly for people with uncontrolled blood sugar levels. When blood glucose levels are high, tiny blood vessels supplying the retina are damaged and leak fluid and blood. A buildup of fluid in the retina can cause it to swell, resulting in vision loss. Preventing the development and progression of DR and preserving vision depends on early detection, proper glycemic control, and timely treatment [31].

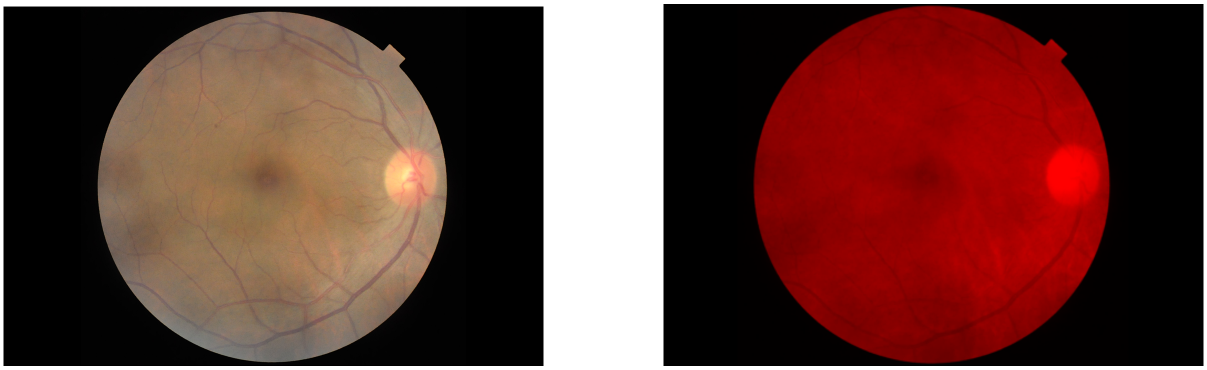
Pathological Myopia (PM): Degenerative myopia, or pathological myopia, is a severe type of nearsightedness that causes structural changes to the eye. Pathological myopia can cause retinal tears because of the thinning of the retina caused by the elongation of the eyeball [31]. It is possible for these complications to cause severe vision loss and, in some cases, even lead to blindness.

Central Retinal Vein Occlusion (CRVO): is a blockage of the central retinal vein, which is responsible for draining blood from the retina. It is common for one eye to suffer a sudden and significant reduction in vision due to obstructions. Hypertension and age-related vascular changes are often associated with CRVO [31].

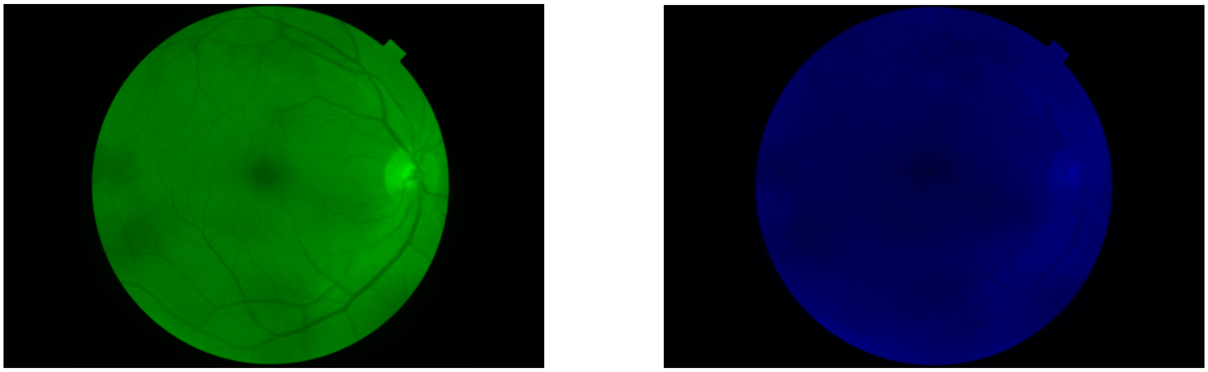
Branch Retinal Vein Occlusion (BRVO): is a vascular disorder that occurs when a branch of the retinal vein becomes blocked, leading to impaired blood flow and tissue damage in the retina [31]. Older individuals are more likely to suffer from the condition, which is often accompanied by medical conditions such as hypertension, atherosclerosis, and diabetes. Sudden loss of vision or blurry vision are common symptoms of BRVO. Depending on how severe the blockage is and where it is located, vision loss varies. We classify central and branch retinal vein occlusion as Retinal Vein Occlusion (RVO).

Glaucoma: is an eye disease caused by high pressure in the eye that damages the optic nerve. Blindness can result if it is left untreated. In order to prevent permanent damage to the optic nerve, early detection through regular eye exams is crucial [31]. Treatment options include eye drops, medication, and surgery.

Age-Related Macular Degeneration (ARMD): is a common eye condition that affects older adults, especially those over the age of 60. The macula, the part of the retina that provides sharp, central vision, gradually deteriorates, resulting in blurred or distorted vision in the center of the field. Two types of ARMD exist: dry and wet, with the latter being more severe. ARMD cannot be cured, but medication, laser therapy, and photodynamic therapy can help slow its progression and preserve vision [31].



(a) Original Image (b) Red Channel



(c) Green Channel (d) Blue Channel

Figure 1.1: Different channels of a fundus image. The Green channel has more invaluable information.

1.2.2 Fundus Images Characteristics

Natural images are photographs of the real world and can include a wide variety of subjects, such as landscapes, people, and objects. In contrast, fundus images are photographs of the interior surface of the eye, including the retina, optic disc, and blood vessels. They are typically taken using specialized cameras that can capture detailed images of the fundus.

One key difference between fundus images and natural images is the level of detail that is important. In fundus images, it is crucial to be able to see small structures and details within the eye, such as the blood vessels and the optic disc. This is because these structures can provide important information about the health of the eye and the presence of certain conditions, such as diabetes or hypertension. In contrast, natural images may not require such a high level of detail and may be more focused on the overall shape and composition of the image.

Another difference between fundus images and natural images is the importance of the color channels. In natural images, all three color channels (red, green, and blue) are typically important for creating a full and accurate representation of the scene. In fundus images, however, the green channel is often more important and contains more valuable data (see Figure 1.1). This is because the blood vessels and other structures within the eye tend to be more visible in the green channel, making it easier to identify and analyze these structures [38]. As a result, many fundus image analysis algorithms place a greater emphasis on the green channel when processing and interpreting the images.

1.2.3 Clinical Challenges

One major challenge in data gathering for fundus image studies is the limited availability of annotated images. Fundus images are often collected as part of routine medical exams and are not always labeled with relevant clinical information. This can make it difficult to obtain large, representative datasets for training and evaluating machine learning models.

Another challenge is the variability in the appearance of fundus images, which can be affected by factors such as the patient's age, the type and severity of the condition being imaged, and the imaging device used. This variability can make it difficult to develop machine learning models that are able to accurately classify the images, particularly if the training data is not representative of the full range of variation in the images.

Finally, data privacy is a major concern in the context of fundus image studies. Fundus images often contain sensitive personal information and medical history, which must be protected in accordance with relevant laws and regulations. This can make it difficult to share and use fundus image datasets for research and development and can limit the availability of data for training and evaluating machine learning models.

1.3 Our Solution

One of the most powerful machine learning models is Deep Neural Networks (DNNs). They are capable of recognizing visual objects and faces, segmenting images, and processing natural language. These models have become ubiquitous in various fields of study, including medical imaging, due to their exceptional performance. One advantage of using pre-trained deep learning networks for solving the challenges of data gathering and data privacy in fundus image studies is the transferability of these networks. Transferability means the ability to acquire and reuse knowledge. Deep neural networks are able to learn complex, high-level representations of the data, which can be transferred to different tasks and domains.

This means that a pre-trained network trained on a large, diverse dataset of images can be used as a starting point for developing a machine-learning model for fundus image classification. The model can then be fine-tuned on a smaller dataset of fundus images, allowing it to adapt to the specific characteristics of the task at hand. Using pre-trained deep learning networks in this way can provide several benefits. First, it can reduce the amount of data and labeling effort required, as the model can be trained on a small dataset of fundus images and fine-tuned using the pre-trained network. Second, it can improve the performance of the model, as the pre-trained network provides a strong foundation of knowledge that can be leveraged to solve the specific task at hand. Finally, it can improve the generalizability of the model, as it can be trained on a wide range of data and then fine-tuned to adapt to the specific characteristics of the fundus images.

The usage of pre-trained deep learning networks can provide a valuable solution to the challenges of data gathering and data privacy in fundus image studies. By leveraging the transferability of these networks, we can develop machine learning models that are accurate, reliable, and generalizable without requiring access to large, labeled datasets of fundus images.

1.3.1 Unsupervised Domain Adaptation

To improve the effectiveness of machine learning models, annotated data is essential. However, the absence of labeled data poses a significant challenge in many real-world scenarios for training these models. The amount of labeled training data available directly affects the performance of machine learning models. The issue is made worse by the time-consuming and expensive process of manual data annotation. One potential solution is transferring knowledge from a labeled domain to a similar but different domain with limited or no labels. Nonetheless, data bias or domain shift can make this approach challenging, as machine learning models often struggle to generalize from an existing domain to an unlabeled domain [54].

Typically, conventional machine learning techniques assume that training and test data come from identical distributions and that models are optimized using training data before being applied to test data. As a result, any differences between training and test data are ignored. Nevertheless, it is a typical scenario that there exist disparities between the source and target domains, and the conventional methodology underperforms in the presence of a domain shift complication. Domain shift occurs when there is a discrepancy between the training and testing data distributions.

By utilizing the knowledge gained from a labeled source domain, domain adaptation (DA) aims to create reliable predictors for a target domain with limited or no labeled data, while addressing domain shift issues.. A DA can be supervised, semi-supervised, or unsupervised, depending on the number of labels in the target domain. Supervised DA provides all target data labels, whereas semi-supervised DA provides only some labels. There are no labels on the target data in unsupervised domain adaptation (UDA). UDA, also known as closed-set domain adaptation, involves an equal number of categories in the source and target domains [54].

The UDA task is the main focus of this study. The extracting of features from raw images was a key component of earlier DA methods. In recent years, researchers have been using high-performance deep neural network features like AlexNet [22], ResNet50 [17], and Xception [9] instead of low-level SURF [5] features.

Source-Free Domain Adaptation. is a specific type of unsupervised domain adaptation where the model is not provided with any information about the source domain, where the training data comes from. This is in contrast to traditional unsupervised domain adaptation, where the model is provided with some information about the source domain, such as the distribution of the data or the features used to represent the data. One advantage of source-free unsupervised domain adaptation is that it allows the model to adapt to a new domain without any assumptions about the source domain. This can be useful in situations where the source and target domains are very different or where the information about the source domain is not reliable or not available.

1.3.2 Open-Set Recognition

Open-Set Recognition (OSR) is a type of machine learning algorithm that is able to recognize and classify items in a dataset that it has been trained on, as well as identify items that it has not seen before. This is in contrast to closed-set recognition, where the algorithm can only classify items within a predefined set of classes. Open-set recognition is useful in situations where there may be new or unexpected items that need to be classified and is often used in applications such as security and fraud detection.

One of the key challenges in OSR is to determine whether a given item belongs to a known class or is a new, unknown item. To do this, the algorithm must have some way of measuring the degree of similarity between an input item and the items in its training dataset. This can be done using various techniques, such as comparing the feature vectors of the input item to the feature vectors of the known items, or using a generative model to estimate the likelihood that the input item belongs to a known class. Once the algorithm has determined that an item is unknown, it can either reject it outright or attempt to classify it into a new, unknown class.

One of the main advantages of open-set recognition is that it allows the algorithm to adapt to new, previously unseen items. This can be especially useful in dynamic environments where the classes and characteristics of items may change over time. For example, in a security application, open-set recognition can help the algorithm identify and classify new types of fraudulent activity that it has not seen before. Another advantage of OSR is that it can help to reduce the risk of misclassification. In a closed-set system, an item that does not belong to any of the predefined classes may be misclassified into one of the known classes, leading to incorrect or misleading results. In an open-set system, the algorithm can explicitly reject items that do not belong to any of the known classes, reducing the likelihood of misclassification. Briefly, OSR is an important advance in the field of machine learning and is an active area of research. It has the potential to improve the performance and flexibility of machine learning algorithms, and to enable them to adapt and learn in real-world environments.

Open-set recognition is particularly important for fundus image classification because it can help to improve the performance of the model on new and unseen data. Fundus images can vary greatly in appearance, and can be affected by factors such as the patient's age, the type and severity of the condition being imaged, and the imaging device used. This variability can make it difficult for a machine learning model to accurately classify the images, particularly if it has not seen examples of the novel classes before. By using open-set recognition, the model can learn to identify and classify novel classes of disease, which can help to improve its overall performance on the fundus images.

1.3.3 Open-Set Domain Adaptation

Open-Set Domain Adaptation (OSDA) is a type of machine learning algorithm that is designed to adapt a model trained on one set of data to a different set of data, where the data in the new domain may contain classes that are not present in the original training data. This is in contrast to traditional domain adaptation algorithms, which assume that the classes in the new domain are a subset of the classes in the original training data.

One key challenge in OSDA is that the model must be able to identify classes in the new data that it has not seen before and must be able to adapt to them without being able to access any additional training data. This requires techniques such as novelty detection and one-class classification, which can help the model identify and classify novel classes in the new data. Another challenge in OSDA is that the model must be able to adapt accurately to the new data while still maintaining the ability to classify the original classes with high accuracy. This can be difficult because the model may be biased towards the original training data and may not be able to generalize well to the new data. To address this issue, researchers have proposed various approaches, including adversarial domain training, which uses an adversarial network to help the model learn domain-invariant features, and multi-task learning, which trains the model on multiple tasks simultaneously to improve its ability to adapt to new data.

Adapting to new data without additional training data is important for fundus image classification, as it enables machine learning models to learn from new data. This is particularly important in the context of medical applications, where data privacy concerns may prevent the use of large, publicly available datasets for training and evaluation.

OSDA can help address these concerns by enabling smaller, private datasets for training and evaluation while still allowing the model to adapt to new data. As a result, the model can generalize well to new images and conditions while protecting the privacy of patients whose images are used for training and evaluation. Furthermore, OSDA can prevent machine learning models from overfitting to specific datasets or conditions, which can be a major concern in medical applications. By allowing the model to adapt to new data without access to additional training data, OSDA can help to improve the model's ability to classify images from a wide range of patients and imaging devices, which can be critical for ensuring the accuracy and reliability of the model in real-world settings.

1.4 Contributions

Our main contributions to open-set source free domain adaptation of fundus images are summarized as follows:

- To better capture the fine-grained details of fundus images, an integrated module of spatial attention and channel attention is utilized. This module extracts both local and channel-wise features, taking into account the unique characteristics of fundus images.
- During the adaptation phase, we introduce a new objective measure for selecting the optimal source model by utilizing spatial and channel-wise information. This approach can effectively handle interclass variations between samples in the target domain.
- Through extensive experiments, we utilized explainable AI to demonstrate our proposed method's effectiveness. As far as we know, this is the first attempt to address open-set domain adaptation of fundus images using a source-free approach.

1.5 Thesis Outline

This thesis outline is organized as follows.

- Chapter 2 introduces the background of unsupervised domain adaptation. A discussion of some existing models for treating inter-domain differences is presented. In chapter 2, we talk about different source-free and open-set domain adaptation solutions in the literature. In addition, we explore the development of deep learning methods for fundus image classification
- Chapter 3 describes the structure of our proposed model in detail. We analyze all parts in detail as well. As part of this chapter, we will explain the reasoning behind the most significant parts of our proposed model.

- Chapter 4 discusses performance evaluation results.
- Chapter 5 summarizes this thesis and discusses possible research directions in the future.

Chapter 2

Background and Literature Review

This chapter presents relevant studies from various areas of study. In section 2.1, we introduce the basic knowledge of unsupervised domain adaptation (UDA). Furthermore, we discuss source-free domain adaptation in section 2.2. Then, in Section 2.3, we describe the open-set domain adaptation (OSDA) methods. Next, we discuss the clinical side of our research on fundus image definition and different methods for fundus image classification in 2.4. Finally, section 2.5 reviews the literature on attentions in deep neural networks.

2.1 Domain Adaptation

Currently, most machine learning models require a large amount of labeled training data in order to be highly effective. Real-world applications cannot satisfy such a requirement. Data annotation is time-consuming and costly due to the limited number of labels. Labeling unlabeled data manually becomes very tedious and a bottleneck in the development of a new domain. In many cases, knowledge must be transferred from one labeled domain to another. As a result, the model performance degrades due to domain shift (differences between domains) [11]. Figure 2.1 shows the domain adaptation when there is a domain shift between source and target distribution.

Domain in machine learning refers to a distribution of data characterized by the joint probability distribution of input features and output labels. Domains are sets of data samples that share similar statistical properties. These properties may include the range and distribution of the feature values, the distribution of the label values, and the relationships between the features and labels. Domain adaptation involves adapting a model trained on one domain (the source domain) to perform well on another domain (the target domain), where the two domains may have different statistical properties. By leveraging the knowledge gained from the source domain, the model is expected to perform better on the target domain. This can be useful, for example, when the target domain has limited labeled data, but a large amount of labeled data is available in the source domain.

An example of a domain in the context of image classification would be images taken under different lighting conditions or images captured from different perspectives. Domains can represent different types of text, such as news articles, social media posts, and scientific publications in natural language processing. Images acquired with different retinal imaging devices are an example of a domain in fundus image analysis. Various fundus cameras use different illumination, lens, and sensor configurations, resulting in varying images. The variations can affect the appearance of retinal structures such as blood vessels, optic discs, and maculas. In Unsupervised Domain Adaptation (UDA), source domains are labeled, and target domains are unlabeled. In UDA, the primary goal is to decrease domain discrepancies between labeled source data and unlabeled target data while learning representations that are domain independent.

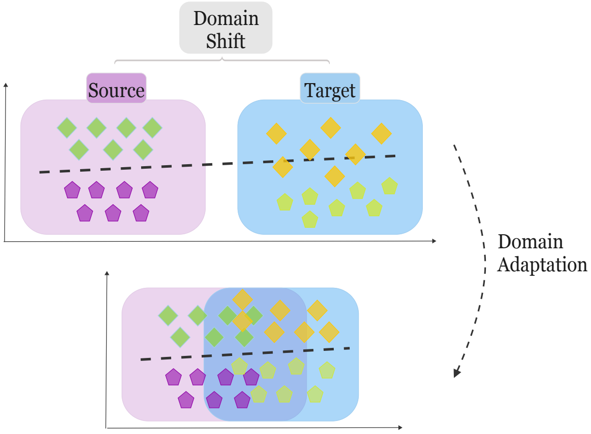


Figure 2.1: Domain adaptation in the presence of domain shift.

Notation. A “domain” consists of a set of features $\mathcal{X}$; for instance, a collection of images, $\mathcal{X} = \{X^j\}_{j=1}^n$. A more complicated kind of domain also includes a set of labels, $\mathcal{Y}$; for instance, some information attached to each member of a subset of $\mathcal{X}$, say $\mathcal{Y} = \{Y^k\}_{k=1}^c$ where c is the number of different classes in the source domain. In Unsupervised Domain Adaptation (UDA), a “target domain” is an example of a minimal domain: it consists only of a set of images without identifying information about their labels. We use $\mathcal{D}_\mathcal{T} = \{X_t^1, \ldots, X_t^{n_t}\}$ to represent a target domain. By contrast, a “source domain” contains much more information: it consists of a set of images, each of which has its own label. We use $\mathcal{D}_\mathcal{S} = (\{X_\mathcal{S}^1, \ldots, X_\mathcal{S}^{n_s}\}, \{Y_\mathcal{S}^1, \ldots, Y_\mathcal{S}^{n_s}\})$. To generalize a model (source model) well to new, unseen data in the target domain, unsupervised domain adaptation transfers knowledge from the source domain to the target domain. The model is trained on labeled data in the source domain and then adapted to unlabeled data in the target domain. A good feature representation is learned from the source domain, which is needed to capture underlying patterns in the data, whereas adapting the model to new, unseen data is a challenge faced in the target domain. Despite the lack of labeled data in the target domain, the model can generalize well to it based on the similarities between the source and target domains. The source samples $\mathcal{X}_\mathcal{S}$ and target samples $\mathcal{X}_\mathcal{T}$ follow the marginal distribution of $P(\mathcal{X}_\mathcal{S})$ and $P(\mathcal{X}_\mathcal{T})$, respectively. The conditional distributions of the two domains are represented by $P(\mathcal{X}_\mathcal{S} \mid \mathcal{Y}_\mathcal{S})$ and $P(\mathcal{X}_\mathcal{T} \mid \mathcal{Y}_\mathcal{T})$, respectively. As the distributions are assumed to be different between the two domains, i.e., $P(\mathcal{X}_\mathcal{S}) \neq P(\mathcal{X}_\mathcal{T})$ and $P(\mathcal{X}_\mathcal{S} \mid \mathcal{Y}_\mathcal{S}) \neq P(\mathcal{X}_\mathcal{T} \mid \mathcal{Y}_\mathcal{T})$, UDA is intended to reduce the gap between domains and train a classifier that exhibits lower generalization error in the target domain.

The literature on domain adaptation categorizes the different types of shifts that can occur between the source and target domains into four main classes: covariate shift, label shift, concept shift, and conditional shift [54]. **Covariate Shift:** In this class, the source and target domains have the same class labels, but the distribution of input features (covariates) differs between the two domains ($P(\mathcal{X}_\mathcal{S}) \neq P(\mathcal{X}_\mathcal{T})$). In adaptation, the input features are mapped to the target distribution by learning a mapping from the source to the target. **Label Shift:** In this class, the source and target domains have different class label distributions ($P(\mathcal{Y}_\mathcal{S}) \neq P(\mathcal{Y}_\mathcal{T})$), which can lead to biased training. The goal of adaptation is to correct this bias by reweighting the source data or by selecting a subset of the source data that is representative of the target domain. **Conditional Shift:** In this class, the conditional distribution of the output given the input is different between the source and target domains ($P(\mathcal{X}_\mathcal{S} \mid \mathcal{Y}_\mathcal{S}) \neq P(\mathcal{X}_\mathcal{T} \mid \mathcal{Y}_\mathcal{T})$). The goal of adaptation is to learn a model that can adjust the conditional distribution to match the target domain. **Concept Shift:** In this class, the underlying concepts or relationships between the input and output differ between the source and target domains ($P(\mathcal{Y}_\mathcal{S} \mid \mathcal{X}_\mathcal{S}) \neq P(\mathcal{Y}_\mathcal{T} \mid \mathcal{X}_\mathcal{T})$). In adaptation, the goal is to learn a model that captures the relevant concepts in the target field.

In spite of several successes of the existing DA models, it remains challenging to minimize domain differences. This section reviews recent domain adaptation papers and introduces a taxonomy based on methods published on UDA.

2.1.1 Domain Adversarial Training of Neural Network

The first adversarial-based DA method is the Domain Adversarial Neural Network (DANN) [12]. In DANN, a gradient reversal layer is integrated with a minimax loss to enhance discrimination between source and target domains. Gradient reversal, which multiplies the gradient by a particular negative constant during backpropagation-based training, prevents the distribution of features across domains from being distinguished. DANN uses a feed-forward neural network to extract features and classify labels. A domain discriminator is added via a gradient reverse layer after feature extraction. As the network trains, a label predictor is minimized for labeled data from the source domain. The network continuously minimizes the domain classifier's loss across all data. There are two weighted components in the optimized objective function of DANN: classifier loss in the source domain and discriminator loss in the target domain.

One way to implement domain adversarial training is through the use of Generative Adversarial Networks (GANs), as proposed by Goodfellow et al. in 2014 [16]. In a GAN, the model generates synthetic data similar to the target domain, while the discriminator is trained to distinguish between real and synthetic data. By training the model to generate data that the discriminator cannot distinguish from the real data, the model learns to generalize to the target domain.

Several unsupervised domain adaptation methods use the GAN structure to solve the problem. Cycle-consistent adversarial domain adaptation [18], which uses a GAN to transfer the style of the source domain data to the target domain data. This is done by training the GAN to generate synthetic data that is both similar to the source domain and consistent with the target domain. The model is then trained using both the real and synthetic data, allowing it to adapt to the target domain. Another approach is Multi-adversarial domain adaptation [33], which uses multiple GANs to learn multiple transformations between the source and target domains. This allows the model to learn a more comprehensive representation of the target domain, leading to improved performance.

2.1.2 Pseudo-labeling in Domain Adaptation

Pseudo-labeling involves generating pseudo-labels based on predicted class probabilities. Classifiers generate pseudo labels for a target domain based on data from a source domain [53]. Pseudo-labels can be used as real labels after they are generated. Utilizing the pseudo-labeling, Saito et al. [36] proposed the asymmetric tri-training structure. In this structure, two classifiers are trained using labeled data from the source domain. Data in the target domain is labeled using the trained classifier in the source domain. Labeling is considered reliable only when both models predict the same label or if at least one classifier predicts a result that is greater than a predefined threshold. Using the pseudo-labeled target domain data, a new classifier is trained to represent the target discriminatively.

Using pseudo-labeling at the batch level is another method of utilizing pseudo-labeling. Chang et al. [8], propose a method for unsupervised domain adaptation that combines pseudo-labeling with domain-specific batch normalization (DSBN) to improve the model's performance on the target domain. Pseudo-labeling is used in the self-training stage of the DSBN method. The model is first trained on the source domain data in this stage using a supervised learning approach. The model is then applied to the target domain data, and its predictions are used as pseudo-labels to retrain the model. This allows the model to adapt to the target domain using its own predictions, leading to improved performance on the task.

2.1.3 Universal Domain Adaptation

Universal Domain Adaptation is a form of domain adaptation that does not demand prior knowledge of label sets [49]. The source label set and target label set may have a shared label set as well as unique label sets, which creates an additional category gap. In Universal Domain Adaptation, the model must either (1) correctly classify the target sample if it corresponds to a label in the shared label set or (2) designate it as ”unknown” otherwise. As the word “universal” implies, universal domain adaptation does not impose any prior knowledge on label sets. Domain adaptation models in the wild face two major technical challenges due to universal domain adaptation. (1) Determining which parts of the source domain are compatible with which parts of the target domain is impossible due to the lack of knowledge about the target label set. Naively matching an entire source domain with an entire target domain will lead to a model weakened by the mismatch between label sets. (2) In cases where the target samples do not belong to any class in the label set, the model must designate them as ”unknown.” Without labeled training data, the classifier cannot determine its specific category. Any approach for universal domain adaptation must have a mechanism for identifying the shared label set.

In the context of universal domain adaptation, You et al. propose [49] Universal Adaptation Network (UAN). UAN has a feature extractor, an adversarial domain discriminator, a non-adversarial domain discriminator and a label classifier. The feature extractor receives input from either domain and extracts the relevant features, which are then passed on to the label classifier to generate the probability of the input belonging to the source classes. Meanwhile, the non-adversarial domain discriminator evaluates the input's domain similarity, measuring its likeness to the source domain. The goal of the domain discriminator is to align source and target feature distributions through an adversarial approach.

2.2 Source-Free Domain Adaptation

To align source features with target features, unsupervised domain adaptation methods need to access the source data. It is possible, however, that raw source data may not be available in many cases, such as medical records, due to the privacy policy. Source-Free Domain Adaptation (SFDA) attempts to overcome this challenge by using trained models rather than raw data from the source domain as supervision and obtains surprisingly effective results. Our next section discusses source-free solutions for unsupervised domain adaptation.

2.2.1 Entropy Minimization and Self-Supervised Learning

Entropy minimization is a method used in self-supervised learning to improve the ability of a model to adapt to new data. This is typically done by training the model on many unlabeled data from the source domain and then using the learned representations to adapt the model to the target domain. The entropy of a model's predictions is a measure of their uncertainty. By minimizing the entropy of the model's predictions, it becomes more certain of its predictions, which can help improve its performance on the target domain. This approach has been shown to be effective in a number of studies and is a good solution for SFDA because it allows the model to learn generalizable representations from the source domain that can be applied to the target domain. This approach has been shown to be effective in a number of studies, and is a promising solution for SFDA. Here we explain how entropy minimization works in the literature and how self-supervised learning can improve results.

SHOT. In Source Hypothesis Transfer (SHOT) [26], the domain adaptation problem is considered as a source-free problem for the first time. The previous UDA methods needed access to the source data when learning to adapt the models. Using both information maximization and self-supervised pseudo-labeling, SHOT learns a target-specific feature extraction module implicitly. To this end SHOT aligns representatives from the target domain to the source hypothesis by freezing the source model's classifier module. This approach assumes that the same deep neural network model is used across domains and consists of a feature encoding module and a classifier module (hypothesis). The SHOT approach involves developing a feature encoding module that is specific to the target domain. This module generates representations of target data that are closely aligned with the representations of the source data, without requiring access to either the source data or the target labels. The fundamental principle behind SHOT is that to achieve source-like representations for target data, the output of the source classifier for the target data should resemble that of the source data, indicating a high degree of similarity, and approaching one-hot encoding.

The SHOT approach involves freezing the source hypothesis and fine-tuning the source encoding module by maximizing mutual information between intermediate feature representations and classifier outputs. This process encourages the network to encode target features with one-hot encodings that are diverse and disparate. However, even with information maximization, there is still a possibility that target feature representations may align with the wrong source hypothesis. To prevent this, Liang et al. [26] proposed a self-supervised pseudo-labeling technique that augments the target representation learning. This involves generating intermediate class-wise prototypes for the target domain and supervising these prototypes to obtain cleaner pseudo-labels. This approach is used to guide the mapping module learning, as the pseudo-labels generated by a source classifier can be noisy and inaccurate for target data.

In the first stage of training on the labeled source domain, SHOT minimizes the cross-entropy loss with the label smoothing technique. In the second stage of training on the target domain, SHOT utilizes information maximization loss based on the intuition that target features are aligned to source features when their classification results from the source classifier are similar. Then there exists a diversity prompting loss for avoiding the trivial solution that all samples being pushed to a few classes; also, there is a self-training, which uses pseudo-labels to supervise the information maximization process.

SHOT++. There are two major extensions to SHOT in SHOT++ [27]. A further self-supervision objective is proposed for SHOT to predict the relative rotation, which helps the model to learn semantically meaningful representations. Furthermore, the authors suggested a labeling transfer strategy, which only requires labeling predictions in the domain of the target. To learn semantically meaningful representations, rotation prediction is added as a self-supervised task. For an image in the target domain, randomly sample an integral number that corresponds to the rotation degree pool $[0°, 90°, 180°, 270°]$.

Labeling transfer. According to the authors, when examining the confidence scores of SHOT predictions using an entropy function ($\mathbb{H}(p) = \sum_i p_i \log p_i$), some less confident (high-entropy) predictions may be inaccurate. In addition, by using valid labeling details from high-confidence predictions, a less-confident prediction can be improved. Through a two-step process, SHOT++ passes information from low-entropy predictions to high-entropy predictions. The target domain is divided into two splits, one labeled subset, and one unlabeled subset, based on the confidence scores. As a next step, SHOT++ employs MixMatch ([6]) to learn improved predictions for the unlabeled set.

2.2.2 Generative Adversarial Network domain Domain Adaptation

In the context of SFDA, GANs can be used to learn generalizable representations from the source domain that can be applied to the target domain. This is typically done by training the generator on a large amount of unlabeled data from the source domain, and then using the learned representations to adapt the model to the target domain. The classification task in a convolutional neural network (CNN) involves predicting a class label for an input image. Convolutions and pooling operations are layered into multiple layers before one or more fully connected layers perform classification. Training involves updating the weights of the network during backpropagation to minimize the difference between predicted and true class labels so the network learns to map input images to corresponding class labels.

CNN discriminators, however, are types of neural networks used in generative adversarial networks (GANs). Based on an input image, the discriminator predicts whether it belongs to the real or fake data distribution. In other words, it is trained to differentiate between real images and synthetic images generated by a generator network. During training, the generator network produces realistic images to fool the discriminator into believing they are real.

The pretrained model in a GAN style can adapt to new data without needing labeled data from the target domain by learning generalizable representations from the source domain. This approach has been shown to be effective in a number of studies and is a promising solution for SFDA. We explore GAN-based methods for SFDA in the following sections.

Universal Source-Free Domain Adaptation

We discussed universal domain adaptation in part 2.1.3. Source-free universal domain adaptation avoids using source samples in the adaptation process. To deal with this issue, universal source-free domain adaptation [23] proposes a two-stage approach. In the first stage (Procurement), assuming no prior knowledge of changing categories and domains, the model will be prepared for future source-free deployment. A generative classifier framework improves the rejection of out-of-source distribution samples. In the second phase (Deployment), the goal is to develop an adaptive algorithm that can be used without access to previous sample data.

In the procurement stage, they propose combining the source images to synthesize hypothetical negative classes using an image composition method that generates new negative samples by combining the positive samples, which can serve to represent unforeseen categories. Synthetic samples are more representative of the expected characteristics in the deployed environment than samples from unrelated datasets. In the development stage, they define the source similarity metric to determine how similar the target samples are to the source samples. The metric's higher value signifies a greater similarity towards the positive source categories and is specifically inclined toward the common label space. Conversely, a lower value indicates a similarity towards the negative source categories. Using the development output, we are able to distinguish between target samples belonging to the shared label set and those belonging to the private label set.

In order to perform domain adaptation, the objective function moves the target samples with higher source similarity metrics in the direction of the positive source collections (from the procurement stage) and vice-versa. They also use entropy minimization ($\mathbb{H}(p) = \sum_i p_i \log p_i$) to move the target samples toward highly confident areas within the classifier's feature space.

2.3 Open-Set Domain Adaptation

As we discussed in section 2.1, domain adaptation aims to train a classifier in a label-rich domain (source domain) and apply it to a label-scarce domain (target domain). A classifier trained in a different domain performs less accurately on samples from different domains due to their distinct characteristics. However, most domain adaptation approaches assume a closed-set assumption, where targets belong to the same classes as sources. This assumption is not always realistic, especially in unsupervised domain adaptation, where only unlabeled target samples are available. Without labels, it is impossible to determine whether the target samples belong to the source domain class. Hence, open-set recognition algorithms are necessary for domain adaptation. Open-set domain adaptation deals with this problem, in which samples in the target domain don't belong to the source domain's class. As shown in Figure 2.2 in open-set domain adaptation, unknown target samples should be classified as "unknown", and known target samples should be classified into their correct categories. As part of open-set domain adaptation, we avoid negative transfer, which means we should not transfer open-set samples between source and target domains.

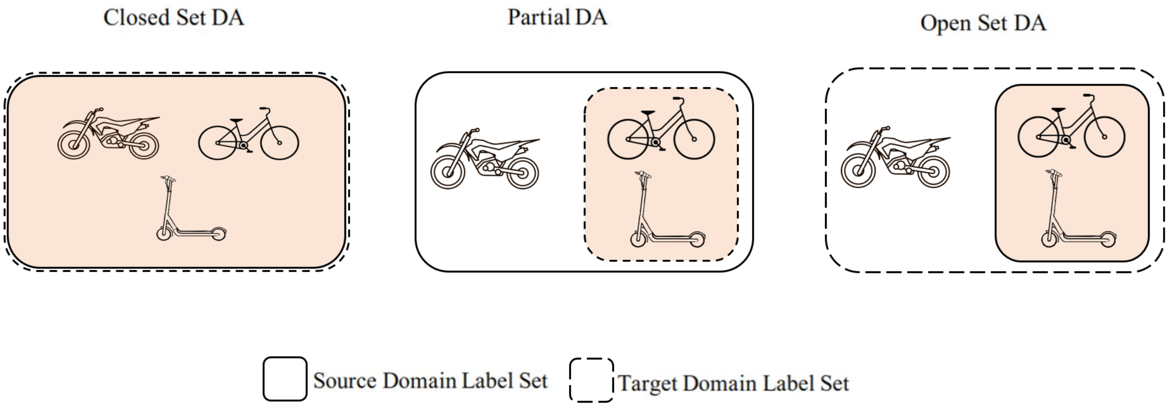


Figure 2.2: Domain adaptation settings based on source and target label sets (Shared labels are indicated by colored rectangles). The image is redrawn from [49]. Copyright © 2019, IEEE.

The first problem in this situation is that we don't know which samples are unknown. Therefore, drawing a boundary between known and unknown classes seems difficult. Domain differences are the second problem. To reduce this domain difference, we must align target samples with source samples. However, unknown target samples cannot be aligned because there are no unknown samples in the source domain. The following section reviews the literature for proposed solutions to open-set domain adaptation.

2.3.1 Instance-Level Approaches

Instance-level approaches align the distributions at the instance level by using techniques such as re-weighting or matching-based methods. These approaches typically aim to identify corresponding instances in the source and target domains and use these correspondences to align the distributions. These approaches can be effective when the source and target domains have a similar structure but may not be as effective when the domains are significantly different.

Backpropagation-based

Dealing with open-set domain adaptation problems, Open-Set Domain Adaptation by Backpropagation (OSBP) [37] proposes a method for facilitating the rejection of unknown target samples and the alignment of known target samples with known source samples. The classifier and the feature generator are two key players in this method. Features are generated from inputs using the feature generator, while the classifier uses these features to produce a probability output of $K + 1$ dimensions, where K represents the number of known classes. The probability of the unknown class is represented by the $(K + 1)$th dimension of the output. Feature generators are trained to distance target samples from decision boundaries, while classifiers are trained to distinguish between source and target samples. In particular, the classifier is trained to output a probability of t for the unknown class, where $0 < t < 1$.

Suppose we have a classifier that is poorly trained and requires improvement to construct an effective decision boundary for unknown samples. In this scenario, the feature generator has two viable options to manipulate the classifier output probabilities. The first option is to align the unknown samples with the source domain, while the second option is to reject them as unknown altogether. By training the classifier and generator on a cross-entropy loss, we can accurately categorize the source samples and improve their performance. A binary cross entropy loss can be used to train a classifier for making boundaries for unknown samples.

Multiple Classifiers

Shermin et al. [40] extend the adversarial model from OSBP and propose an adversarial domain adaptation model with multiple auxiliary classifiers. As part of the proposed multi-classifier structure, a weighting module evaluates distinctive domain characteristics to assign weights to target samples that are more representative of whether they belong to known or unknown classes. It simultaneously encourages positive (shared classes between source and target) transfers during adversarial training and reduces the gap between source and target domain classes while at the same time reducing the domain gap between them. To avoid negative transfers, the authors analyze the discriminative domain information of known and unknown target samples and allocate weights to them based on their similarities with the source domain. For this purpose, they introduce a weighting module that assesses each target sample and generates a weight that represents the underlying discriminative domain information. The module assigns distinct weights to known and unknown target samples by comparing their similarity with the source domain, aiding the generator in determining whether to decrease or increase the probability of the ”unknown” class and align the target samples with known or ”unknown” classes.

Progressive Separation

The Separate to Adapt algorithm (STA) [29] gradually separates samples of known and unknown classes while simultaneously weighing the alignment between each class and the feature distribution. STA focuses on considering the effect of negative sampling, which means aligning the entire distribution of source and target domains would be risky since unknown classes in the target domain can further lower the domain adaptation model’s performance than a model without adaptation. This requires accurate identification of the boundary between known and unknown classes, even without knowing anything about the unknown classes. Adaptation should also be applied to the known classes in both domains.

2.3.2 Hybrid Approaches

Hybrid approaches combine multiple methods to achieve domain adaptation. Diverse methods can be more flexible and effective than traditional approaches when the source and target domains differ significantly.

Self-Supervision Using Roatation

Rotation-based Open-Set (ROS) [7] is a two-stage approach. In the initial phase, ROS distinguishes between the known and unknown target samples by building a model that predicts the degree of rotation between a reference image and its rotated version. In order to minimize the discrepancy between the source domain and the known target domain, the authors repeat the rotation task once more. The last step is to obtain a classifier capable of categorizing each target sample into a known class or disregarding it as an unknown sample.

Clustering-Based

Based on global image features, Domain Consensus Clustering (DCC) [25] proposes a method for performing category-level clustering for OSDA. DCC proposes cycle-consistent matching to associate common cluster centers (i.e. common classes) across domains. The pair of clusters that are the nearest centers to each other in different domains is considered a common cluster, whereas the unmatched clusters are rejected as unknown outliers. Further, it optimizes the number of clusters searched by computing the sample-level consensus and promotes the effectiveness of cycle-consistent matching. Specifically, they draw the domain consensus knowledge from two aspects to facilitate clustering and the discovery of private classes, which include a domain-level understanding that recognizes common clusters as the common classes and a sample-level consensus that determines clusters and private classes using cross-domain classification agreements.

Another clustering method proposed Pan et al. [32]. They first apply clustering to the source domain to identify category-agnostic clusters. These clusters are then used to weight the contributions of different domain adaptation techniques. The resulting domain-adapted model is then applied to the target domain and is able to handle the presence of new categories by assigning them to the appropriate category-agnostic cluster. The main idea behind this approach is to use clustering to identify corresponding instances in the source and target domains, but rather than assuming that these clusters correspond to specific categories, the approach allows for the possibility of new categories in the target domain. One advantage of this approach is that it allows for the possibility of new categories in the target domain, while still making use of the structure in the source domain to guide the domain adaptation process. This makes it more flexible than approaches that assume that the categories in the target domain are a subset of the categories in the source domain.

Source-Free Open-set Domain Adaptation

Authors in [24] present a method called Inheritable Models for Open-Set Domain Adaptation (IMOSDA) for addressing the problem of open-set and source-free domain adaptation. In this problem, the target domain includes both seen and unseen classes, and the goal is to adapt a model trained on a source domain to perform well on the target domain. IMOSDA addresses this problem by introducing a transferability score that measures how well a model can adapt to the target domain. This score is calculated using a transferability network, which is trained to predict the transferability of each feature in the source model to the target domain. The transferability score is used to guide the adaptation process and ensure that the adapted model is able to recognize both seen and unseen classes in the target domain. To encourage the model to learn transferable features that are shared between the source and target domains, IMOSDA introduces an objective called the inheritability loss. This loss encourages the model to learn features that are not only discriminative for the seen classes in the target domain, but also transferable to the unseen classes.

2.4 Fundus Image Analysis

We will discuss the clinical components of our research in the following section. Also, we will explore the use of various deep learning-based models for the classification of fundus images. Deep learning has recently gained significant attention in the field of medical image analysis due to its ability to learn complex patterns and features directly from data. We will investigate the performance of different deep learning architectures on fundus image classification tasks in the next section.

2.4.1 Deep Learning Ensemble Approach

Diabetes Retinopathy (DR) damages the retinal blood vessels. Without early diagnosis, DR can lead to blindness and impaired vision. DR can be five stages or classes: normal, mild, moderate, severe, and PDR (Proliferative Diabetic Retinopathy)[14]. Highly trained professionals examine colored fundus images to diagnose this fatal disease. In this case, clinicians have to make a manual diagnosis that is time-consuming and error-prone. The detection of DR from retinal images has therefore been proposed using a variety of computer vision-based techniques. However, it is important to note that these methods cannot encode the complex underpinnings of DR, meaning they can only classify the various stages of DR with a very low degree of accuracy, particularly in the early stages. Qummar et al. [34], train an ensemble of five deep Convolution Neural Network (CNN) models (Resnet50 [17], Inceptionv3 [41], Xception [9], Dense121 [20], Dense169 [20]) capturing rich features and improving classification through DR. Furthermore, light networks have been studied for their ability to reduce convolutional network complexity. Gayathree et al. [13] present a convolutional neural network architecture to extract features from retinal fundus images to classify DR in binary and multiclass manners. The proposed method reduces complexity while improving classification accuracy. CNN extracts symbolic information from the input data and makes layer-by-layer abstraction possible through the layer-by-layer stacking of convolution, pooling, and non-linear activation function mapping. Their major contribution is significantly reducing the CNN model parameters to enable real-time deployment while improving classification accuracy.

Furthermore, another ensemble network is proposed in [45]. There are two parts to the model: the first is a feature extractor, which includes a transfer learning-based model that consists of a pre-trained model with no top layer. A multi-label classifier based on the above features makes predictions in the second part. Combining two weak classification models creates a more robust classification model. As part of the integration strategy, the original and grayscale images are first equalized by histograms. Two identical training data sets are used for the operation. Both sets of similar data sets are independently trained using the EfficientNet [42] model. Finally, both models are averaged based on their sigmoid output probabilities. The authors opted for EfficientNet because it achieved superior accuracy compared to other models while utilizing significantly fewer parameters and floating-point operations per second (FLOPS) on both ImageNet [10] and five widely used transfer learning datasets.

2.4.2 Transformers for Fundus Image Classification

Multiple Instance Learning Enhanced Vision Transformer (MIL-VT) [50] uses an extensive fundus image database to train the vision transformer model and then refine downstream tasks for retinal disease classification. The authors suggest utilizing a ”MIL head” based on multiple instance learning (MIL) that can be easily added to the vision transformer to improve the performance of downstream fundus image classification models. The proposed framework outperforms CNN models when trained and tested under the same conditions, as demonstrated using two publicly available datasets. In MIL, an image is regarded as a bag containing pixels or image patches, which is similar to the relationship between image patches and the vision transformer in the paper. To fully utilize the features of individual patches, the authors employ multi-instance learning to the vision transformer structures. This involves creating low-dimensional embeddings for ViT features from individual patches, using an aggregation function to obtain the bag representation, and applying a bag-level classifier to obtain the final bag-level probabilities.

2.4.3 Graph Neural Network for Fundus Image Classification

Lin et al. [28] propose two novel multi-label classification networks to classify fundus images; Multi Classification Network based on Graph convolutional networks (MCG-Net) utilizes graph convolutional networks, while Multi Classification Network based on Graph convolutional networks and Self-supervised learning (MCGS-Net) incorporates both graph convolutional networks and self-supervised learning. The graph convolutional network is designed to extract the relevant information from multi-label fundus images, and self-supervised learning is employed to improve the network’s generalization capability. In order to better capture the correlation between fundus images, they construct MCG-Net using a graph convolutional network (GCN) instead of a fully connected layer as a classifier. By adding a module for self-supervised learning (SSL) to MCG-Net, they built MCGS-Net, that improved MCS-Net’s generalization ability. MCGS-Net improves classification performance by learning more unannotated fundus images using the generalization enhancement module. MCGS-Net consists of three components: the backbone module, C_{GCN} module (GCN for Classification), and G_{SSL} module (SSL for Generalization). The backbone module is used for sharing feature extraction. The C_{GCN} module is used for multi-label classification. The G_{SSL} module is designed for the generalization enhancement of MCGS-Net. For the SSL part, the network learns image feature information by predicting the transformation type of the fundus image. As well as this, the authors define two types of geometric transformation classification based on rotations of 0 or 90 degrees. SSL usually employs unannotated data and a pretext task for training and then uses the target dataset for fine-tuning.

2.4.4 Domain Adaptation for Fundus Images

The open-set domain adaptation of fundus images is addressed for the first time in [56]. Zhou et al. present a Collaborative Regional Clustering and Alignment (CRCA) method for identifying category-agnostic local feature clusters. By using a cluster-based method they match common local regions rather than identifying common classes, which is more fine-grained for learning domain-invariant features. They also use informative region selection that utilizes Class Activation Map (CAM) [55] to compute the importance weight map of each class. By applying CAM to dense local features, computational complexity is reduced. A cluster-aware contrastive adaptation method is proposed to retain the local features of aligned clusters across domains while pushing those of misaligned clusters far away. The framework explores local features to learn domain-invariant representations to guide distribution adaptation. In addition, they propose a benchmark and dataset for OSDA on fundus images.

2.5 Attention

Attention mechanisms in computer vision have their roots in research on human vision and cognitive science. It has been observed that humans only notice a portion of all visible information due to limitations in information processing. Researchers have attempted to model this selective attention process in order to understand how humans distribute their attention when observing images, videos, and other visual stimuli, and to apply this understanding to various fields. In recent years, significant progress has been made in using attention mechanisms for image and natural language processing. It has been shown that attention mechanisms can improve model performance and are also consistent with the perceptual mechanisms of the human brain and eyes. In the field of computer vision, much research on combining deep learning and attention mechanisms focuses on the use of masks, which identify key features in image data through a layer with new weights. Through training, deep neural networks can learn to pay attention to specific areas of each new image, resulting in attention. This idea has evolved into two types of attention: soft attention, which is differentiable and continuous and is realized through gradient descent, and hard attention, which is not differentiable and is often achieved through reinforced learning with a reward function that encourages the model to pay more attention to certain details. As attention mechanisms in computer vision have evolved, different models have emerged that pay attention to different feature domains. This section will provide examples of some of these models.

2.5.1 Soft attention

The differentiability of soft attention has made it a popular tool in various computer vision applications, including but not limited to classification, detection, segmentation, model generation, and video processing. Soft attention mechanisms can be classified into several categories, including spatial attention, channel attention, mixed attention, and self-attention [50].

Spatial attention

Convolutional neural networks (CNNs) can exhibit translation and implicit rotation invariance in their learning. However, an explicit processing module may be more effective for handling these transformations compared to networks that learn them implicitly. Consequently, DeepMind designed Spatial Transformer Layer (STL) to realize spatial invariance [21].

The localization network initially computes a θ value for the input image U. The grid generator then utilizes this value along with the coordinates of the output image to calculate the corresponding coordinates of the input image. Finally, the sampler fills in the output image V based on predefined rules of filling (usually bilinear interpolation is employed). By utilizing these steps, the input image can be rectified into the desired image through spatial transformer learning.

Channel attention

In a convolutional neural network, an image is initially represented by three channels (red, green, and blue). After being processed by different convolution kernels, each channel generates new channels that contain different information. If weights are assigned to each channel to reflect their relevance to key information, a higher weight indicates higher relevancy and the corresponding channel should receive more attention.

SENet, the winner of the ImageNet [10] Classification Contest in 2017, is essentially a channel-based attention model [19]. It models the importance of each feature channel and then enhances or suppresses it in different tasks. After the normal convolution, a bypass branch is applied, which involves squeezing the spatial dimension features, resulting in each 2D feature map being compressed to a single real number. The subsequent step involves excitation, where a weight ”w” is generated for each feature channel to explicitly model the relevance. Once the weight of each feature channel is determined, it is applied to each original feature channel, enabling the learning of the significance of different channels based on specific tasks. By adding a relatively small amount of computations, the channel attention mechanism can lead to notable enhancements in performance across various benchmark models. Moreover, this mechanism has broad applicability, as it can be integrated into many existing networks. For example, SKNet incorporates channel weighting into the multi-branch network structure of inception to achieve improved results. Essentially, the channel attention mechanism can model the significance of different features and customize the weights accordingly to suit the specific task. This approach is straightforward yet impactful.

Self attention

In a convolutional neural network, the size of the convolution kernel limits its ability to access local information for calculating the target pixel, which can result in errors due to the lack of global context. By treating each pixel in the feature map as a random variable and calculating the pairwise covariances, the value of each predicted pixel can be adjusted based on its similarity to other pixels in the image. This process of using similar pixels in training and prediction and ignoring dissimilar pixels is known as the self-attention mechanism. In order to achieve global reference for each pixel-level prediction, a non-local Neural Network using self-attention in CNN was proposed by [46]. Their method considers each pixel as a random variable based on the predicted covariance between pixels. The participating target pixels are then obtained by taking the weighted sum of all pixel values, where each pixel's weight is determined by its correlation with the target pixel. By utilizing the self-attention mechanism, the model can achieve global reference during both training and prediction, resulting in a more reasonable bias-variance weight.

2.5.2 Hard attention

The soft attention mechanism has been widely and successfully applied in the field of computer vision. In contrast, research on the hard attention mechanism in computer vision tasks is more limited. Hard attention is seen as a more efficient and direct approach because it can select important features from input information. While the role of constraints such as sparsity in shaping the learning ability of agents has been explored, AttentionAgent took a different approach and was inspired by concepts related to inattentional blindness, which is the phenomenon of the brain focusing most attention on elements related to a task and temporarily ignoring other signals while engaged in a task requiring effort.

2.6 Summary

This chapter provided a comprehensive overview of various techniques for domain adaptation, namely unsupervised domain adaptation, source-free domain adaptation, and open-set domain adaptation. Additionally, we delved into the use of deep learning techniques for fundus image classification. Moreover, attention mechanisms in deep networks were discussed, highlighting their crucial role in boosting the performance of neural networks. Understanding these topics is fundamental to grasp the current scenario of domain adaptation and fundus image classification, as well as the significance of incorporating attention mechanisms in deep learning models.

Chapter 3

Methodology

In this chapter, we will focus on our source-free Open-Set Domain Adaptation (OSDA) where, in the absence of the source dataset, the source model is utilized to facilitate adaptation to open-set unlabeled data by delving into channel-wise and local features for fundus disease recognition. In particular, considering the nature of the fundus images, we present a novel objective way in the adaptation phase to utilize spatial and channel-wise information to enable the selection of the most suitable source model for a given target domain, even by considering the small inter-class variation between samples.

3.1 Problem Formulation

Suppose $n_{\mathcal{S}}$ labeled images $\mathcal{D}_{\mathcal{S}} = (\{X_{\mathcal{S}}^{1}, \ldots, X_{\mathcal{S}}^{n_s}\}, \{Y_{\mathcal{S}}^{1}, \ldots, Y_{\mathcal{S}}^{n_s}\})$ are drawn from a source density $P(\mathcal{X}_{\mathcal{S}}, \mathcal{Y}_{\mathcal{S}})$ and $n_{\mathcal{T}}$ unlabeled images $\mathcal{D}_{\mathcal{T}} = \left(\mathbf{x}_i^{\mathcal{T}}\right)_{i=1}^{n_{\mathcal{T}}}$. $C_{\mathcal{S}}$ indicates the set of the source classes, $C_{\mathcal{T}}$ shows that of the target, and $C_{\mathcal{T}}/C_{\mathcal{S}}$ denotes the implicit classes in the target domain. In OSDA, due to $C_{\mathcal{T}}/C_{\mathcal{S}} \neq \emptyset$, we are required to classify target samples of $|C_{\mathcal{S}}|$ known classes correctly ($|A|$ indicates the number of members in A) and concurrently drop the unknown target samples belonging to $C_{\mathcal{T}}/C_{\mathcal{S}}$. In the OSDA setting, additional *unknown* classes C_{unk} only exist in the target domain label space ($C_{\mathcal{T}} = C_{\mathcal{S}} \cup C_{unk}$), and make up of $(N + 1)$ classes in total. The objective is to label each instance in the target set by assigning a class for the shared classes $C_{\mathcal{S}}$ and an *unknown* label for the unshared classes (C_{unk}). The model trained on the source domain is denoted as M_s, while the model adapted to the target domain is denoted as M_t.

3.2 Method Overview

As shown in Figure 3.1, the architecture of OSDA includes two phases: (1) training the source model M_s on the source dataset $\mathcal{D}_{\mathcal{S}}$ in an open-set setting as shown in the top of the figure. (2) adapting the target model M_t to the target dataset $\mathcal{D}_{\mathcal{T}}$ given by the trained source model M_s as illustrated in the bottom of the figure. During phase 1, we synthesized negative samples as unknown classes and trained the source model on a combined dataset from source dataset $\mathcal{D}_{\mathcal{S}}$ and the unknown classes. In phase 2, we employ the inherit-tune paradigm mentioned in chapter 2 as the main flow of adaptation and plug our proposed Spatial and Channel-wise

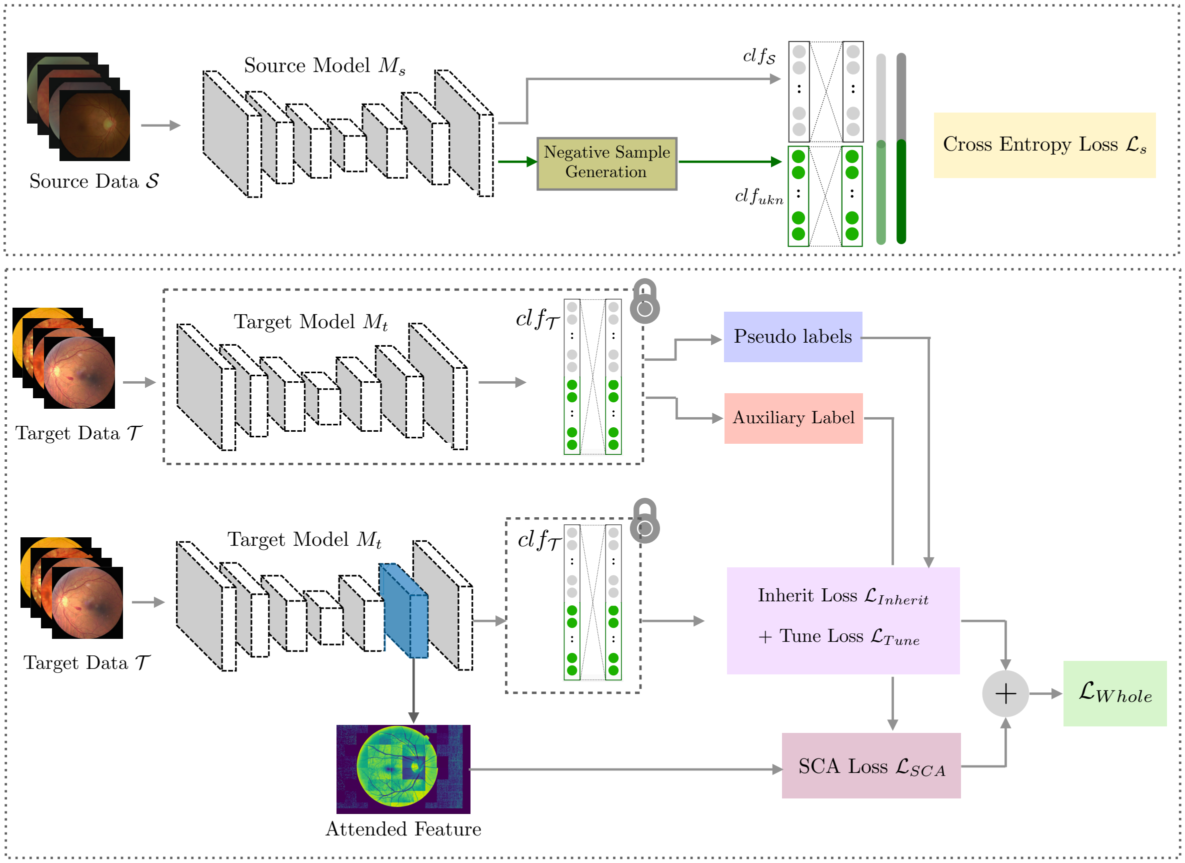

Source Data $\mathcal{S}$
Source Model M_s
$clf_{\mathcal{S}}$
Negative Sample Generation
clf_{ukn}
Cross Entropy Loss $\mathcal{L}_s$
Target Data $\mathcal{T}$
Target Model M_t
$clf_{\mathcal{T}}$
Pseudo labels
Auxiliary Label
Target Data $\mathcal{T}$
Target Model M_t
$clf_{\mathcal{T}}$
Inherit Loss $\mathcal{L}_{Inherit}$
+ Tune Loss $\mathcal{L}_{Tune}$
$\mathcal{L}_{Whole}$
SCA Loss $\mathcal{L}_{SCA}$
Attended Feature

Figure 3.1: Overview of our method.

Adaptation (SCA) component into it to take full advantage of the fundus images' nature. In Section 3.3, we describe the process of the source model training in an open-set setting. In Section 3.4, we will focus on the target model adaptation which includes the process of Inherit, Tune, and our SCA respectively.

3.3 Source model training

We apply negative sample generation to train the source model in an open-set setting. In particular, we adopt the background-class-based modeling approach to solve the problem of unknown samples in neural networks by adding new classes as representative of unknown samples during training. Despite the fact that there are a number of works investigating the generation of negative samples, for simplicity, In order to swap out the top-d percentile activations of a specific feature layer with the corresponding activations from an instance of a different class, we utilize the feature-splicing technique[43, 24, 54] as equation 3.1.

$$u_n = \phi_d\left(u_{\mathcal{S}}^{c_i}, u_{\mathcal{S}}^{c_j}\right) \text{ for } c_i, c_j \in \mathcal{C}_{\mathcal{S}}, c_i \neq c_j \tag{3.1}$$

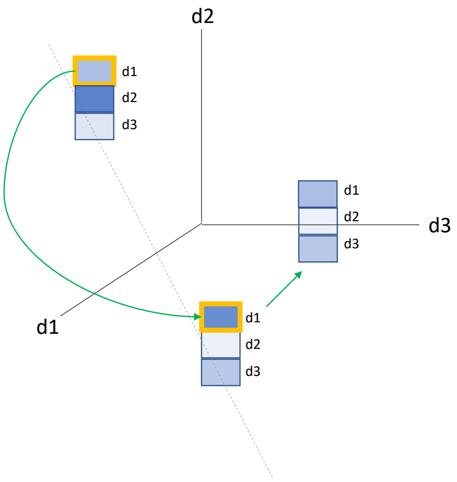


Figure 3.2: We apply feature-splicing by eliminating the class-discriminative characteristics, which involves replacing the activations in the top-(1/3) percentile (d1). The image is redrawn from [24]. Copyright © 2020, IEEE.

In this context, $u_{\mathcal{S}}^{c_i} = M_s\left(x_{\mathcal{S}}^{c_i}\right)$ is obtained by applying the function M_s to the source image $x_{\mathcal{S}}^{c_i}$, which belongs to class c_i in the source data. The feature-splicing operator ϕ_d is then used to replace the top-d percentile activations in the feature $u_{\mathcal{S}}^{c_i}$ with the corresponding activations from $u_{\mathcal{S}}^{c_j}$, where c_j is a different class, as illustrated in Figure 3.2. This results in a feature that lacks class-specific characteristics but is close to its source distribution. To classify these negative samples, we utilize K-means clustering and give each cluster a unique label, as described in [43, 24]. We refer to the unknown samples as $\mathcal{U} = \{(u_n, y_n)\}$. Additionally, in the next chapter, we will present an ablation study where we examine other methods we investigated for generating negative samples.

To train the source model, we first pre-train using source data $\mathcal{S}$ by taking the standard cross-entropy loss as below:

$$\mathcal{L}_{pre} = \mathcal{L}_{CE}\left(\sigma\left(Clf_{\mathcal{S}}\left(M_s\left(x_{\mathcal{S}}\right)\right)\right), y_{\mathcal{S}}\right) \tag{3.2}$$

The term $Clf_{\mathcal{S}}$ refers to the classification layer in the source model, as depicted in Figure 3.1. In addition, σ denotes the softmax activation function. Afterward, we train the source model using a combined dataset from $\mathcal{S}$ and $\mathcal{U}$ as follows:

$$\mathcal{L}_s = \mathcal{L}_{CE}\left(\hat{y}_{\mathcal{S}}, y_{\mathcal{S}}\right) + \mathcal{L}_{CE}\left(\hat{y}_{\mathcal{U}}, y_{\mathcal{U}}\right) \tag{3.3}$$

where $\hat{y}_{\mathcal{S}}$ comes from nodes belonging to source classes in the classification layer $Clf_{\mathcal{S}}$ and $\hat{y}_{\mathcal{U}}$ denote nodes belonging to unknown classes in the classification layer Clf_{unk}.

In this case, using $\hat{y}_{\mathcal{S}}$ model can extract the class-separability knowledge and $\hat{y}_{\mathcal{U}}$ can be used to construct an understanding of negative samples.

3.4 Target model adaptation

To adapt the target dataset using the unlabeled target dataset, we begin by initializing the target model with the trained source model. Adaptation involves three processes, namely Inherit, Tune, and Spatial and Channel-wise Adaptation (SCA), where the first two are based on the basic inherit-tune paradigm [24] and the third process is our novel adaptation objective. We first review the structure of Inherit and Tune and then deep dive into spatial and channel-wise adaptation processes.

3.4.1 Inherit

The Inherit process is intuitively intended to facilitate class separation. In particular, we can use the source model's uncertainty for an input sample to find out how inheritable the model is. In detail, the classification confidence obtained from the softmax layer of the source model M_s can measure inheritability as follows:

$$w(x) = \max_{c_i \in C_s} \left[\sigma\left(Clf_s\left(M_s(x)\right)\right)\right]_{c_i} \tag{3.4}$$

The softmax activation function σ is applied to the entire output of Clf. However, it is important to note that the maximum value is computed only over the classes learned from the source dataset, and not the unknown classes. To enable the model to inherit the characteristics of the entire target dataset $\mathcal{D}_{\mathcal{T}}$, additional steps are required. we can define model inheritability $\mathcal{I}$ as follows:

$$\mathcal{I}\left(M_s, \mathcal{S}, \mathcal{T}\right) = \frac{\text{mean}_{x_{\mathcal{T}} \in \mathcal{T}}\, w\left(x_{\mathcal{T}}\right)}{\text{mean}_{x_{\mathcal{S}} \in \mathcal{S}}\, w\left(x_{\mathcal{S}}\right)} \tag{3.5}$$

For a given triplet $\{M_s, \mathcal{S}, \mathcal{T}\}$, a constant $\mathcal{I}$ is assigned, with higher values indicating smaller domain-shift and an increased ability to inherit knowledge. This results in a class separation in the open-set domain adaptations. To ensure that the Inherit process is equipped with class-separability knowledge, we select the top-k percentile target instances based on their $w\left(x_{\mathcal{T}}\right)$ value, and designate them as samples with pseudo-labels, which we refer to as $\mathcal{P} = \{(x_{\mathcal{P}}, y_{\mathcal{P}})\}$. Note that we consider both $Clf_{\mathcal{S}}$ and Clf_{unk} when calculating $w\left(x_{\mathcal{T}}\right)$. However, we assign each percentile that comes from the unknown classes to the same label as $N + 1$. Intuitively, these top-k percentile target instances are those target samples that the model strongly believes that they belong to source classes. Thus we collect them as samples with pseudo-labels and use them in the Inherit process.

To inherit the knowledge of class separation, we utilize the cross-entropy loss to ensure that the target predictions align with the pseudo-labels, as demonstrated below:

$$\mathcal{L}_{Inherit} = \mathcal{L}_{CE}\left(\sigma\left(Clf_{\mathcal{T}}\left(M_t\left(x_{\mathcal{P}}\right)\right)\right), y_{\mathcal{P}}\right) \tag{3.6}$$

3.4.2 Tune

The tuning process is intuitively intended to minimize the effect of negative transfer. Many studies have investigated entropy minimization as a means of guiding the features of unlabeled instances toward high-confidence regions when label information is unavailable. However, like [24], we utilize a loss formulation that involves soft instance weights. Consequently, target instances with higher values of w are directed towards the feature space of the source data, whereas those with lower values of w are pushed towards the feature space of negative samples. Thus, by using the target classifier, we can estimate how likely an instance is to belong to a shared class as below:

$$\hat{S} = \sum_{c_i \in \mathcal{S}} \left[\sigma\left(Clf_{\mathcal{T}}\left(M_t\left(x_{\mathcal{T}}\right)\right)\right)\right]_{c_i} \tag{3.7}$$

By optimizing the loss function below, we encourage the separation of both shared classes and unknown classes.

$$\mathcal{L}_{t1} = -w\left(x_{\mathcal{T}}\right)\log(\hat{s}) - \left(1 - w\left(x_t\right)\right)\log(1 - \hat{s}) \tag{3.8}$$

We also calculate probability vectors separately for shared classes as $p_t^{sh} = \sigma\left(Clf_{\mathcal{S}}\left(M_t\left(x_{\mathcal{T}}\right)\right)\right)$ and unknown classes $p_t^{unk} = \sigma\left(Clf_{\mathcal{U}}\left(M_t\left(x_{\mathcal{T}}\right)\right)\right)$ and minimize the following loss.

$$\mathcal{L}_{t2} = w\left(x_{\mathcal{T}}\right)\mathrm{H}\left(p_t^{sh}\right) + \left(1 - w\left(x_{\mathcal{T}}\right)\right)\mathrm{H}\left(p_t^{unk}\right) \tag{3.9}$$

The symbol H represents Shannon's entropy, given by the expression $H = -\sum p(x)\log p(x)$. The total loss $\mathcal{L}_{\text{Tune}} = \mathcal{L}_{t1} + \mathcal{L}_{t2}$ matches shared classes in a selective manner, while also avoiding negative transfer.

3.4.3 Spatial and Channel-wise Adaptation (SCA)

The utilized Spatial and Channel-wise Adaptation (SCA) process [44] is intuitively constructed to take advantage of fundus images' channel- and spatial-wise characteristics to maximize the adaptation as a complementary step to the Tune process. SCA utilizes channel and spatial attention to concentrate on critical information and its location in the fundus image, facilitating the differentiation of shared and unknown classes. The Channel and Spatial Attention Module is described in detail, followed by an explanation of its integration in the adaptation process.

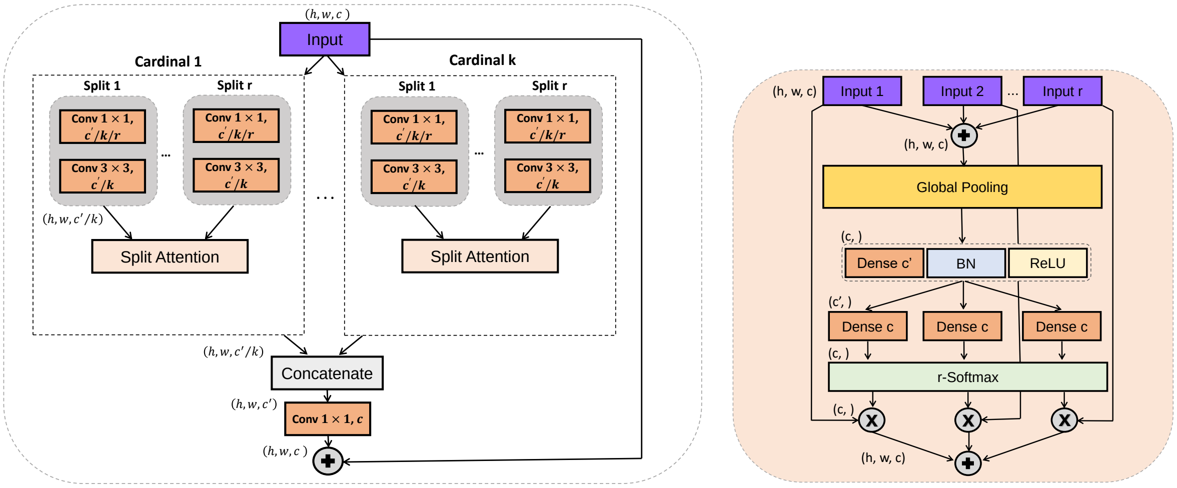
Integrated Spatial Attention and Channel Attention Module (ISCA)

To account for the unique qualities of fundus images, a Spatial Attention (SA) and Channel Attention (CA) module was developed and illustrated in Figure 3.5. In this module, SA is employed alongside ResNeSt's channel-wise attention to emphasize the important information and location of feature maps. Additionally, the module utilizes max pooling to capture the global characteristics of fundus images.

ResNet

To address the problem of vanishing gradients in very deep networks, residual networks (ResNets) [17] were introduced. To link one layer's input to another layer's output, ResNets use skip connections, also called identity mappings. As a result, gradients can propagate more easily through the network and vanishing gradients can be prevented in deep networks.

Using skip connections, ResNets can learn residual functions, which capture the difference between layers' outputs and inputs. Instead of simply passing the input through a sequence of



(a) (b)

Figure 3.3: (a) ResNeSt represents the height, width, and number of channels of the feature map with h, w, and c. In addition, each cardinal group is divided into r splits.(b) The ResNeSt architecture incorporates a split attention module. Image reprinted from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022.

nonlinear transformations, the output of a layer is added to the input. By combining lower-level features from earlier layers, skip connections allow the network to learn higher-level features, which can help improve the network's accuracy.

ResNets were developed as a solution to the problem of vanishing gradients, which can occur in very deep neural networks. The training process may become very slow or even cease when gradients become so small that updating the weights becomes impossible. Through skip connections, ResNets solve this problem by allowing gradients to flow more easily through the network, even in very deep architectures. The skip connections enable the network to learn residual functions, which can capture the difference between the output of a layer and its input.

ResNeXt

As an extension of the ResNet architecture, ResNeXt [48] addresses the issue of overfitting in deep convolutional neural networks and the lack of generalization. Essentially, ResNeXt uses a grouped convolution operation, which divides input channels into groups and applies a separate convolution operation to each group before concatenating the outputs. With this method, the network can learn representations of features that are more diverse and powerful than with standard convolutional neural networks. ResNeXt has an advantage over ResNet in that it can learn more diverse and powerful feature representations, which can improve accuracy and reduce overfitting.

ResNeXt defines cardinality as the number of parallel pathways or groups within a grouped convolution process. By determining the cardinality parameter, we can divide our input channels into groups and apply convolution operations to each group. Increased cardinality allows the network to learn more diverse and powerful feature representations but at the cost of in-

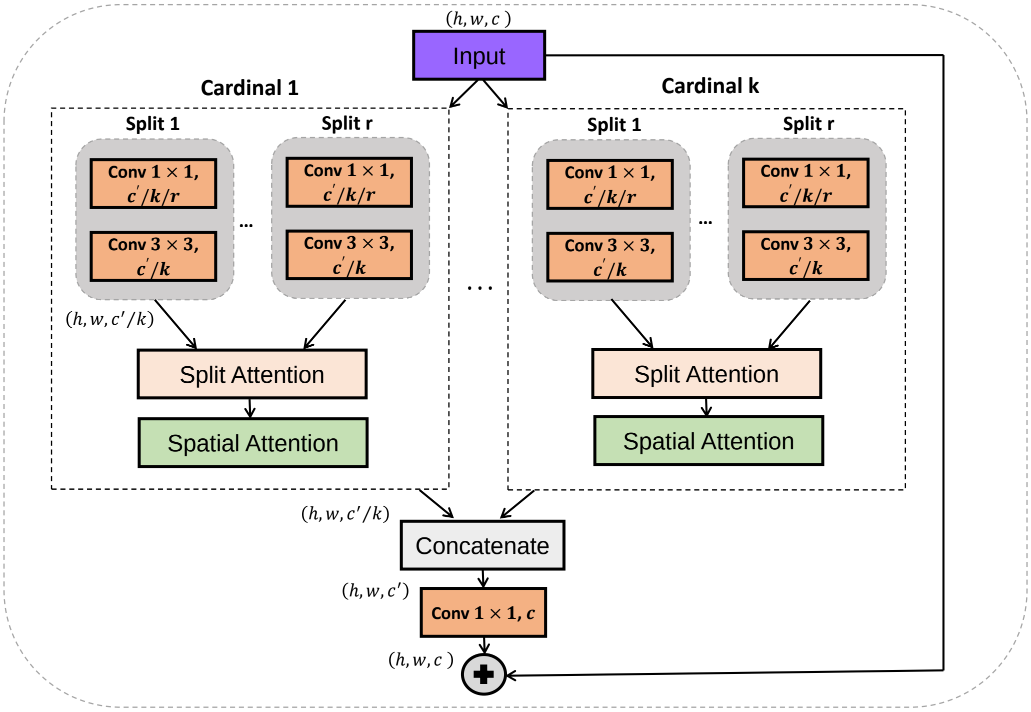


Figure 3.4: An overview of our integrated channel attention and spatial attention module. (The image is reprinted from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022.)

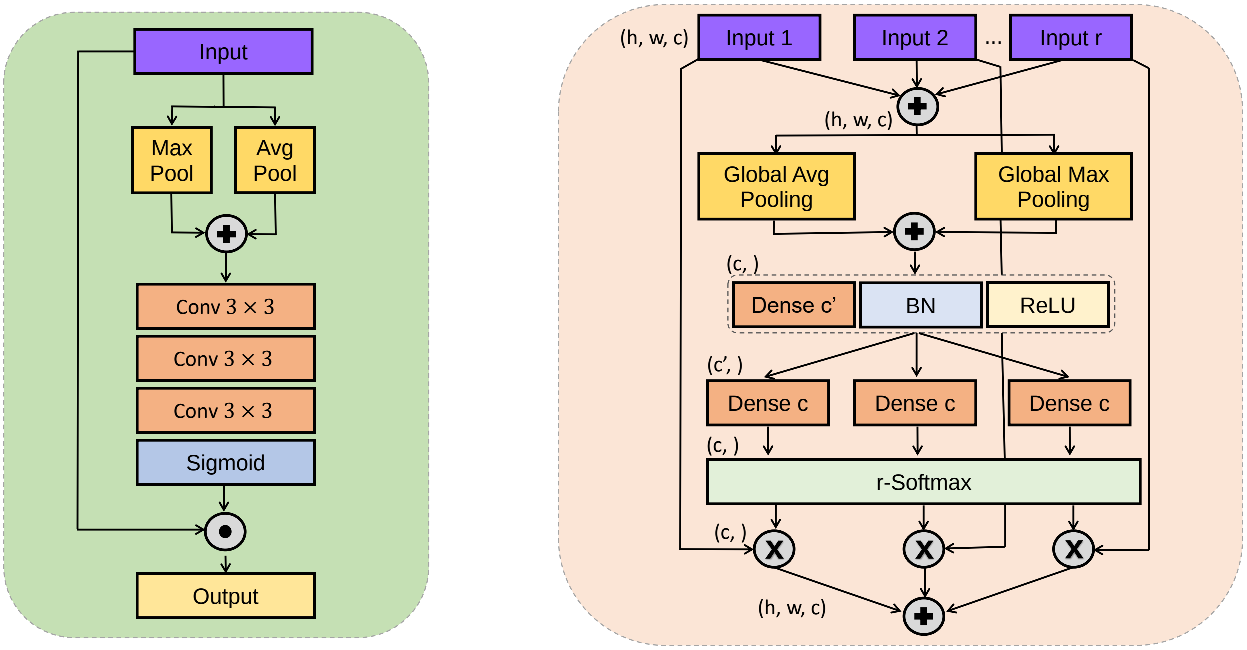
creased computational complexity.

ResNeSt

ResNeSt [52] is a variation of ResNet [17] that features a split-attention block. While retaining the original ResNet framework, ResNeSt incorporates group convolution from ResNeXt [48] and a channel-wise attention mechanism, facilitating the exchange of information among cross-feature map groups and acquiring feature information from different processing regions [44]. A diagram of the ResNeSt block is provided in Figure 3.3.

Channel Attention Module

The utilized CA module is based on the split attention module employed in ResNeSt, illustrated in Figure 3.5.(b). The variety of texture information present in fundus images makes it possible to simplify feature complexity by eliminating redundant and unimportant texture features before calculating weights. To obtain more precise channel-wise attention, a max-pooling operation is employed according to [47]. In the original ResNeSt, global pooling was combined with local pooling to extract contextual information, eliminate noise, and extract texture information more effectively. It can be written as $M(F) = \text{maxPool}(F)$, where F is the input feature map.



(a) (b)

Figure 3.5: (a) Our Spatial Attention (SA). (b) Our Channel Attention (CA). (This figure reproduced from [44]. Licensed under a Creative Commons Attribution (CC BY 4.0), Frontiers, 2022.)

Spatial Attention Module

The relevance of information varies across different positions of fundus images. For instance, retinal blood vessel information in the optic disc tends to be more critical than that in other locations. Thus, it is essential to reinforce such vital information through the use of Spatial Attention(SA). ResNeSt's split attention module utilizes channel-wise attention only to determine feature relationships and importance within channels. Using an integrated attention module, a SA module is incorporated following a Channel Attention(CA) to generate a two-dimensional SA map [44]. By focusing more on spatial positions, SA complements and expands channel-wise attention. By assigning weights to each spatial position, it is possible to identify the most important information about that position while inhibiting extraneous features from being extracted. In order to optimize spatial information, a weight-shared SA block is utilized following channel-wise attention. Figure 3.5.(a) depicts the structure of the SA module.

To consolidate channel information from the feature map, the SA module employs a combination of average-pooling and max-pooling approaches. Following the average and maximum pooling, the output is concatenated and passed through a fully connected layer to generate an attention map. According to this attention map, each channel is assigned a weight based on how relevant it is to the task at hand. Following that, the attention map is used to reweight the feature map, giving more weight to channels that provide more information while suppressing those that don't.

$$M1(F) = [\text{maxPool}(F) \odot \text{avgPool}(F)] \tag{3.10}$$

In the above equation, F represents the input feature map, and the symbol $\odot$ denotes the concatenation operation.

Subsequently, the receptive field of the feature map is broadened suitably via three 3×3 convolution operations. Based on the feature map, a sigmoid function is applied to generate a two-dimensional SA map. In addition, the CA module integrates this information with global information. The calculation can be expressed as follows:

$$M2(F) = \sigma\left(f^{3\times3}\left(f^{3\times3}\left(f^{3\times3}(F)\right)\right)\right) \tag{3.11}$$

Here, the symbol σ represents the sigmoid function, and $f^{3\times3}$ denotes a 3×3 convolution operation.

Finally, the weights from the final SA module are applied to the original feature map using a weighting operation, expressed as:

$$M3(F) = W \times F \tag{3.12}$$

where W represents the weight obtained from SA module [44].

This network architecture can enhance the expressive power of the original image and improve classification accuracy by considering both local and global information.

Spatial and Channel-wise objective

As explained in the Tune process, we desire target instances with higher w points toward the source data feature space, while those with lower w push into the negative sample feature space. In this regard, we employed top-k percentile target instances based on their w in the Tune process as the model strongly believes they belong to the source classes. In the SCA, we argue that the second top-k percentile comes from those samples belonging to the source classes while having some minor variation in details. In this case, we propose that instead of expecting the model to have the same output for them, have intermediary features near to source classes features in the feature space. Thus, we collect the second top-k percentile named auxiliary data and obtain their intermediary features as $\mathcal{A} = \{(x_{\mathcal{A}}, f_{\mathcal{A}}, y_{\mathcal{A}})\}$. In order to obtain the labels $y_{\mathcal{A}}$ for these samples, we use the first top-k $\mathcal{P}$ as a reference. In particular, for $x_{\mathcal{A}} \in \mathcal{A}$ we find the nearest intermediary feature in the corresponding features of $\mathcal{P}$ and assign that label to the sample.

It is important to note that the intermediary features are derived from our proposed Integrated Spatial- and Channel-wise Attention Module. Consequently, the attended features provide practical details and contextual information. By defining a new objective on the spatial- and channel-wise attended parts, we encourage a separation of shared classes and unknown classes while taking into account the nature of fundus images. To fit our adaptation problem in fundus images, we boost the Tune process by optimizing the subsequent loss.

$$\mathcal{L}_{SCA} = \mathcal{L}_{CE}\left(\sigma\left(Clf_{\mathcal{T}}\left(M_t\left(x_{\mathcal{A}}\right)\right)\right), \hat{y}^l_{\mathcal{A}}\right) \tag{3.13}$$

where l is a hyperparameter that selects the layer number of the network from which we get the spatial- and channel-wise attended features. Generally, the early layers of the network model provide fundamental information, whereas the later layers provide more abstract information.

Thus, the final adaptation loss is a summation of the losses obtained by Inherit, Tune, and SCA process as below:

$$\mathcal{L}_{whole} = \mathcal{L}_{Inherit} + \mathcal{L}_{Tune} + \mathcal{L}_{SCA} \tag{3.14}$$

3.4.4 Summary

This chapter concludes by demonstrating that spatial channelwise attention can be used to solve the open-set domain adaptation problem. Using average-pooling and max-pooling, an attention map is generated that assigns weights to each channel based on how relevant it is to the task at hand. Using the proposed spatial channelwise attention-based solution, we can extend the concept to source-free domain adaptation. An adaptation in which the source domain is not available during training is known as source-free domain adaptation, which is more challenging than traditional domain adaptation. As a result, the model must learn to generalize to the target domain without labeled data from the source domain.

Chapter 4

Experimental Analysis

In this chapter, we will present the results obtained from experimenting with various parameter settings. Our model's superior performance compared to other models was demonstrated through analyzing fundus images in an OSDA scenario. Section 4.1 will provide information about the experimental parameters and settings. Additionally, in Section 4.1.2, we will introduce the two datasets we used. The test results of our model will be described in detail in Section 4.2. Finally, we will investigate the reasoning behind the proposed model through explainable methods in Section 4.3.

4.1 Experimental Settings

4.1.1 Implementation Details

Following [24], we implemented our model using PyTorch 1.8 and the proposed structure from the previous chapter was used as the backbone model. During the training of the source model, we used a batch size of 32, with 16 source and 16 negative instances. For the target model, a batch size of 32 was used. All experiments were conducted on a machine with the following hardware specifications: NVIDIA GeForce GTX 1080 GPU with CUDA v11.7. Image augmentations such as horizontal-flip, random rotations, and color jitter were applied during the training process. In the source training phase, the Adam optimizer was used with a learning rate of 1e-4. The training was done for 4000 mini-batches, and the best model was saved based on validation data. A validation set of 15% of the data was used. For the adaptation phase, the Adam optimizer was used with a learning rate of 1e-5. The adaptation was carried out for 15000 mini-batches. For more detailed information about the structure of networks, kernel sizes, and datasets used, please refer to https://github.com/masoudpz/os-sf-da-on-fundus-images.

4.1.2 Datasets

In order to evaluate OSDA methods for fundus disease recognition, we used three datasets (TAOP [3], ODIR [2], and RFMiD [31]) to build two source and target domain pairs. The source domain data is TAOP, contains 3,297 images of five retinal diseases. ODIR and RFMiD, each containing 6,576 and 2,451 images covering more disease categories (covering five classes in the source), are set as target domain data. We call these two open-set fundus image benchmarks OSF-T2O and OSF-T2R. Due to the fact that three datasets were collected from different hospitals and whose images were captured by different fundus cameras, noticeable shifts between domains are apparent. In our analysis, we use 5 classes (Diabetic Retinopathy, Retinal Vein Occlusion, Pathological Myopia, Age-related Macular degeneration, and Glaucoma) as our source data, and other classes as our target data.

ODIR: The ODIR dataset [2] was created for the ”International Competition on Ocular Disease Intelligent Recognition”. This database contains information about 3500 patients, including age and color fundus photographs of both eyes, as well as doctors' diagnostic keywords. The resolution of the fundus images can vary depending on the camera used, such as Canon, Zeiss, and Kowa. The patients are categorized into eight groups based on the provided data, as shown in Figure 4.1, which include normal (N), diabetic retinopathy (D), glaucoma (G), cataract (C), age-related macular degeneration (A), hypertensive retinopathy, and myopia.(H), myopia (M), and other conditions.

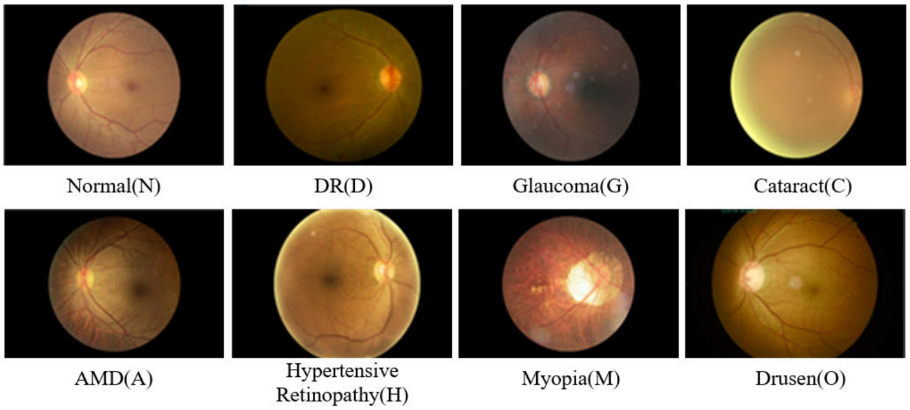


Figure 4.1: ODIR dataset samples and different classes [2].

RFMiD: The Retinal Fundus Multi-Disease Image Dataset (RFMiD) [31] provides an opportunity to detect multiple diseases and develop automated methods to classify frequent and rare eye conditions. Two retinal experts have annotated 46 conditions based on 3200 fundus images captured with three different cameras. RFMiD is categorized based on annotations into two groups: 1) screening for normal and abnormal retinal images, and 2) classification of retinal images into 45 different categories. Samples from various categories in RFMiD are depicted in Figure 4.2.

TAOP: The ophthalmology department of Beijing Tongren Hospital provides clinical samples in the TAOP dataset [3]. This collection comprises 3297 images classified into five categories, which include glaucoma, age-related macular degeneration, diabetic retinopathy, pathological myopia, and retinal vein occlusion. As the source data, we use this dataset in combination with two additional datasets as the target data. Several samples from the TAOP dataset are illustrated in Figure 4.3.

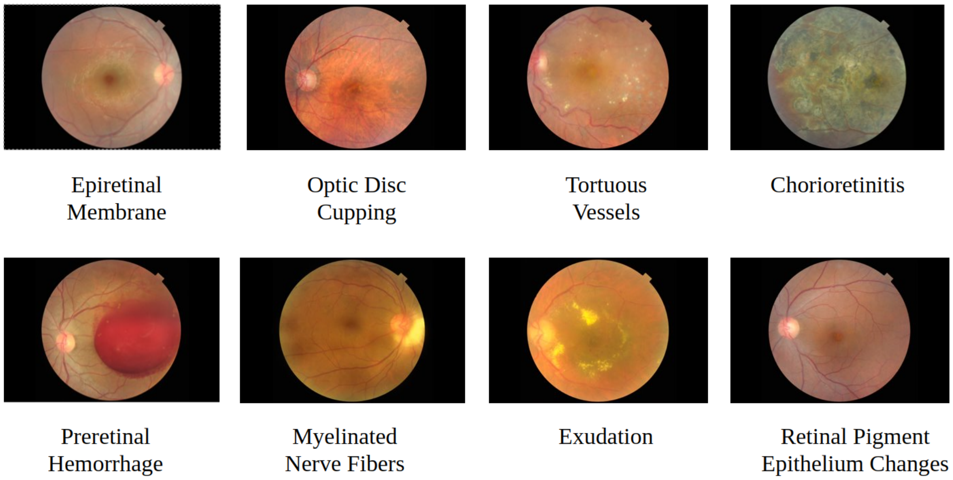


Figure 4.2: RFMiD dataset samples and classes [31].

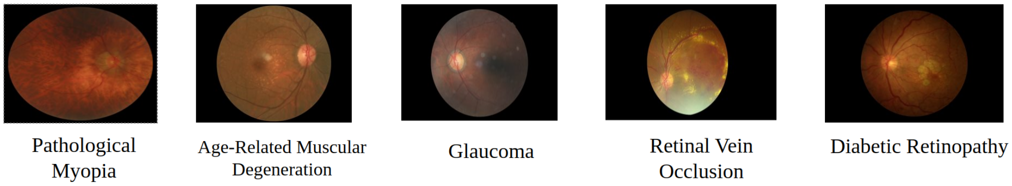


Figure 4.3: TAOP dataset samples and classes [31].

4.1.3 Metrics

Following [56], we evaluate the OSDA performance based on four metrics, i.e. OS*, OS, UNK, and harmonic mean (HM) accuracy. OS* and OS represent the mean accuracy over common classes and the mean accuracy over all classes, respectively. UNK is a measure of accuracy in recognizing unknown classes, and HM is a harmonic mean accuracy as below:

$$\mathbb{HM} = 2 \times \mathrm{OS}^{*} \times \mathrm{Unk} / \left(\mathrm{OS}^{*} + \mathrm{Unk}\right) \tag{4.1}$$

where $\mathbb{HM}$ is a balanced evaluation metric that correctly assesses the performance of the methods on both known and unknown class samples.

4.2 Results

4.2.1 Comparison Methods

DANN [12]. The key component of this method is the use of a gradient reverse layer, which is a type of layer that multiplies the incoming gradient by a negative constant during training. This has the effect of reversing the gradient and allowing the main network to learn features that are invariant to the domain. The gradient reverse layer is used in combination with a domain classification loss term in the overall loss function, which encourages the main network to be domain-invariant while still accurately classifying the examples in the training set.

OSBP [37]. The main part of this method involves training a model on the source domain using backpropagation and then adapting it to the target domain using a combination of backpropagation and an additional loss term that encourages the model to forget the source domain classes that are not present in the target domain. By using this approach, the model can be adapted to the target domain while still being able to classify examples accurately from the target domain classes.

ROS [7]. The main idea investigates the use of image rotation as a way of augmenting the training data for the model in the target domain and shows that this can improve the model's performance when adapting to the target domain. The novelty of the ROS lies in its focus on the use of image rotation as a means of improving open-set domain adaptation.

DAMC [40]. The main is to propose a method for open-set domain adaptation that involves using an adversarial network with multiple classifiers. This method is an extension of the OSBP method, with the addition of multiple classifiers and a new weighting method.

UAN [49]. This model works based on quantifying sample-level transferability to discover the common label set, and the label sets private to each domain. This allows the model to adapt effectively to the automatically discovered common label set while also being able to recognize unknown samples successfully. The method involves using domain-agnostic features, which are designed to be robust to domain shift, and a universal model that can adapt to any target domain.

DCC [25]. The key novelty of the DCC method lies in its use of consensus clustering, which involves training multiple clustering models on the source domain data and then combining the cluster assignments produced by these models to obtain a consensus clustering. This allows the DCC method to learn more robust cluster assignments that are less sensitive to the specific choice of clustering model. The DCC method also involves the use of domain adaptation constraints, which are used to align the feature spaces of the source and target domains, and domain adaptation regularization, which helps to prevent overfitting to the source domain.

CRCA [56]. The main idea of the CRCA method is to identify common local feature patterns that are category-agnostic and then use these patterns to adapt the distributions of the source and target domains. A key component of the CRCA method is the use of a cluster-aware contrastive adaptation loss, which is introduced to adapt the distributions based on the common local features. The contrastive adaptation loss helps to improve the performance of the adapted model by aligning the distributions of the source and target domains in a way that is based on the common local feature patterns.

4.2.2 Quantitative Results

We compare our work to recent methods presented in Table. 4.1. As shown in the table, our model consistently outperforms other methods by clearing gaps in both benchmarks, showing satisfactory improvements. DANN [12], which is a classic closed-set DA method that aligns distributions across domains without trying to separate classes. A simple design makes it ideal for medical images with small inter-class discrepancies. As compared to DANN, OSBP [37] achieves an increase in HM of more than 2%. However, for recent OSDA methods, ROS [7]

Datasets	**OSF-T2O**				**OSF-T2R**			
Methods	**OS***	**OS**	**UNK**	**HM**	**OS***	**OS**	**UNK**	**HM**
DANN[12]	46.859	48.702	57.918	51.805	53.057	53.403	55.132	54.075
OSBP[37]	48.276	50.482	61.510	54.095	52.566	54.519	64.286	57.838
ROS[7]	38.558	40.874	52.456	44.445	41.150	39.196	29.426	34.149
DAMC [40]	45.269	45.256	45.192	45.231	43.067	43.131	43.454	43.259
UAN [49]	47.805	47.923	48.513	48.156	48.846	50.037	55.993	52.176
DCC[25]	38.704	40.289	48.215	42.939	45.244	46.093	50.338	47.656
CRCA [56]	52.538	54.891	65.945	58.483	55.004	57.391	**69.874**	61.554
Ours	**56.482**	**58.835**	**70.6**	**62.75**	**59.057**	**60.641**	68.651	**63.484**

Table 4.1: Result comparison (%) of state-of-the-art OSDA methods

utilizes self-supervision through image rotation to train a binary classifier to identify unknown samples, and DAMC [40] proposes a non-adversarial domain discriminator. There are clear gaps between different classes of natural vision benchmarks, so these methods work well on them. However, performance decreased noticeably for medical images due to the small inter-class discrepancy in global feature spaces. A significant performance drop is experienced by the DCC method [25] due to its poor image-level clustering that focuses on training category-level clusters and separating unknown samples simultaneously. Despite DCC, CRCA [56] proposes a method to separate images of common classes from private classes by clustering and aligning common local features.

4.2.3 Ablation Study

Negative Sample Generation

In Section 3.3, it was stated that an open-set model can address the overconfidence issue when utilized for open-set DA. To achieve this and following the literature in [24], we trained a negative sample classifier Clf_{unk}. Various techniques were considered to generate negative samples. As stated earlier, the feature-splicing technique proved effective in training an inheritable open-set DA model. The adaptation performance of inheritable models, trained using different negative sample generation techniques, on OSF-T2O is reported in Table 4.2. Here, we will discuss another strategy we examined.

Linear interpolation between classes. To obtain a negative feature, we adopt the approach of randomly selecting a pair of source instances that correspond to different classes and then performing linear interpolation between their features. The interpolation is carried out as proposed in [43]. The idea behind this technique is inspired by mixup [43], which interpolates latent features to generate less confident predictions. However, in our experiments (reported in Table 4.2), linear interpolation didn't generate negative samples as effectively as feature-splicing did. The reason for this is that linear interpolation produces features only from a limited range of the source classes, as explained in Section 3.3.

Datasets	**OSF-T2O**			
Methods	**OS***	**OS**	**UNK**	**HM**
Interpolation	48.012	47.517	45.042	46.479
Feature Splicing	56.482	58.835	70.6	62.75

Table 4.2: Adaptation performance (%) of different negative sample generation strategies. OS* and OS represent the mean accuracy over common classes and the mean accuracy over all classes, respectively. UNK is a measure of accuracy in recognizing unknown classes, and HM is a harmonic mean accuracy.The results indicate that feature splicing is more effective than interpolation in generating negative samples.

	OSF-T2O			
Baseline Network	**OS***	**OS**	**UNK**	**HM**
ResNet-50	54.646	56.117	63.472	58.72
ResNeSt-50	55.026	57.671	69.726	61.509
ResNeSt-50 + ICSA	56.482	58.835	70.6	62.75

Table 4.3: Adaptation performance (%) of different backbones and the effect of proposed ISCA module.

Effect of ISCA Module

Table 4.3 shows the performance of three different backbones: ResNet-50, ResNeSt-50, and ResNeSt-50 + ICSA module. The results show that the third backbone, ResNeSt-50 + ICSA, has the best performance. This indicates that the combination of the ResNeSt-50 architecture with the ICSA module is effective at improving the model's ability to extract fine-grained features and details from the dataset, leading to better overall performance. These findings highlight the importance of selecting a backbone architecture that is able to effectively extract the relevant features and details, especially for fundus images, as this can significantly impact the model's performance. It can be concluded that the more the model can extract fine-grained features and details, the better the result is likely to be.

4.3 Model Interpretation

Since deep learning networks are composed of multiple layers and a large number of parameters, they are initially perceived as black boxes. This has led to the development of a field of study called "deep learning interpretability" to explain and interpret these models. We can use interpretation techniques to diagnose network flaws that lead to incorrect predictions when models perform worse than humans. The results can then be used to improve the network and enhance performance by fixing those problems. We must determine which visual features each neuron in a CNN responds to. Several studies have attempted to visualize the type of information coded in each layer of the network and the preferred stimuli of neurons. Initially, these efforts were limited to the first layer and its neurons since it is the layer right after the input

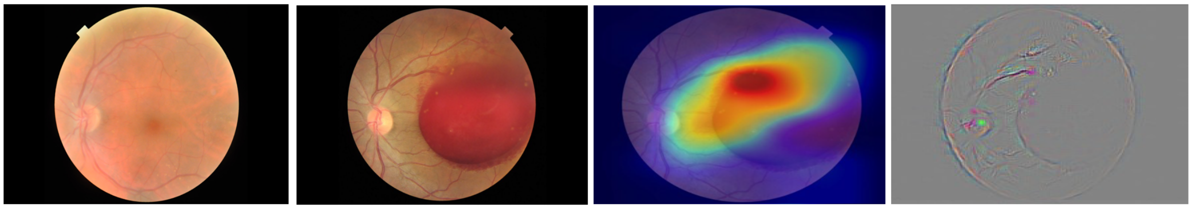


Figure 4.4: Saliency map results of fundus images. From left to right, the normal fundus image, the fundus image with red abnormal hemorrhage, the Grad-CAM, and the Guided backpropagation output. The results obtained from both GradCAM and guided backpropagation demonstrate the excellent ability of the model to locate the place of abnormalities or diseases accurately.

image where inverse projection to pixel space can be made. According to previous studies, the first layer filters are mostly sensitive to basic visual features like edges and colors. A more complex mapping to pixel space is required for higher layers, which cannot be achieved with straightforward mapping. Because the image or feature maps pass through convolutional filters, sampling modules (max pooling), and non-linear activation functions at every layer, the inverse path is challenging since some are irreversible. To visualize higher layers and their neurons, researchers had to develop more complex and creative methods [30].

4.3.1 Saliency Map

To understand why the model predicts a particular label, it is best to determine which parts of the input image played a role. For example, the areas of the image that guide the model's prediction can provide insight into how the model works when used in a clinical application that predicts or diagnoses a disease. Such information can be obtained from saliency maps, called pixel attribution maps or attention maps. Saliency maps are basically images with the same size as the input image, in which each pixel represents how the input image's corresponding pixel contributed to the predicted label. Analyzing the saliency map for multiple dataset instances can reveal what features guide the model's predictions. In addition, incorrectly labeled images can be detected, so the model can't make the correct predictions due to the misleading parts. To this end we investigate our model using two explainable methods.

Guided Backpropagation.

Using deconvolutional networks [51], we can backpropogate the activation of neurons in higher layers to the input space. By inversely implementing each CNN layer, the activation of the desired layer can be used to project the information back to the input space. To see what feature in the image that particular neuron is sensitive to, we can also use one particular filter in that layer instead of using the entire layer and setting the activation of other neurons to zero. Due to the sampling layers, the inverse network itself cannot be reversed in this approach, so we need to start with an input image to be able to use it. Therefore, we need to feed the original CNN with an arbitrary input image (usually a sample from the training set). During the forward path, pixels sampled in max-pooling layers are indexed. During the backward path of the deconvolutional neural network, this information is used to reconstruct the input image. Using the reconstructed image, we can determine what features of the input image were preserved.

Gradient-weighted Class Activation Map (Grad-CAM).

Grad-CAM [39] is another method of interpreting CNN networks based on gradients. However, the gradient does not backpropagate from the label of interest back to the input image. Instead, the gradient is calculated until the last convolutional layer before the first fully connected layer. This gradient represents the contribution of each filter to the classification label of interest in the last convolutional layer. The average of the feature maps is then calculated and weighted by the gradient values. To represent the importance of each region in the input image for the final prediction, this average map should be upsampled to the same size as the input image.
Figs. 4.4 illustrate saliency map results for different samples in the RFMid dataset. The left panel shows the original fundus image, while the middle and right panels display the output of the Grad-CAM algorithm and guided backpropagation, respectively. The model's attention is primarily focused on the optic disc, retinal vasculature, and fovea, as indicated by these visualizations. We can see from Figure 4.4 that the pattern of attention is not uniformly distributed across the entire image. Also, we can see how our model focuses on specific details of each part of the image to detect the type of disease. The proposed network is more effective at diagnosing different fundus diseases due to its ability to focus on details rather than whole shapes. Also, Figure 4.5 shows the effectiveness of the attention module when it comes to detecting different classes. By adding attention, the model could locate more specific features related to the disease. The output of gradcam algorithm is shown in this figure.

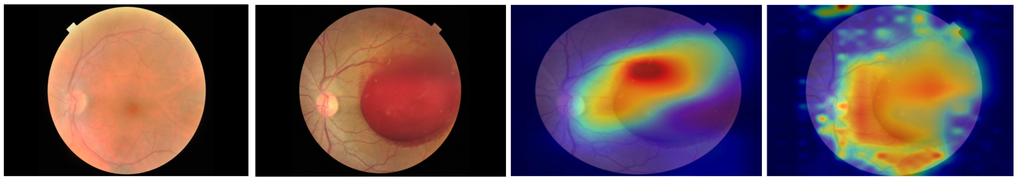


Figure 4.5: Comparing the performance of Resnet and ResNeSt plus attention module on detecting local features. from left to right, the normal fundus image, the fundus image with red abnormal hemorrhage, grad-cam output for Resnet and grad-cam output for the ResNeSt with attention module. ResNeSt, equipped with the proposed module, outperforms the ResNet network in terms of localization of features on fundus images.

4.4 Summary

This chapter provided a detailed account of the implementation process and experiments conducted in our research. We also conducted an ablation study to assess the impact of each component on the overall performance. Additionally, we employed explainable AI techniques to gain insights into the reasons behind our results. Together, this chapter offers a comprehensive summary of the experimental procedure and provides crucial information for replicating and furthering our work.

Chapter 5

Conclusion and Future Directions

5.1 Conclusion

In this thesis, we presented a novel approach for open-set source-free domain adaptation in fundus image analysis. By using convolutional neural networks with spatial and channel-wise attention, we were able to achieve state-of-the-art performance in this field, even outperforming non-source-free methods. Our proposed method is designed specifically for the analysis of fundus images, which can be challenging due to the low inter-class variation between classes. While different classes in general images may have distinct characteristics, in fundus images, classes are often very similar except for small details. Our method addresses this issue and is able to effectively classify and analyze fundus images despite the low inter-class variation.

The process we proposed in this study involves training a source model that is able to generate and identify negative samples and then adapting the target model to the source model using pseudo-labeling and entropy minimization. This approach allows us to effectively adapt to novel domains without access to source data, which is crucial in the open-set context where the number and nature of the target domains are unknown.

In this work, we introduced a novel approach for source-free domain adaptation of fundus images. Our solution to the challenge of capturing fine-grained details combines spatial and channel attention mechanisms to extract both local and channel-wise features. As part of the adaptation phase, we proposed a new objective measure based on spatial and channel-wise information.

One key feature of our proposed method is the use of explainable AI techniques to provide insights into the decision-making process of the model. By using these techniques, we are able to better understand the features and patterns that the model is using to make predictions, which can be especially useful for fundus image analysis where the differences between classes may be subtle. By providing these explanations, we can gain a deeper understanding of the performance of the model and identify areas where it may be possible to improve its accuracy. In addition, the use of explainable AI methods can help to increase the transparency and accountability of the model, which can be critical in the healthcare context where an accurate and reliable diagnosis is critical.

The results of our experiments demonstrate the effectiveness of our method in improving the performance of fundus image analysis tasks in unseen domains. The use of attention mechanisms in our model allows for the preservation of important details in the images, which is critical for the accurate analysis of fundus images. This work makes a valuable contribution to the field of domain adaptation and has the potential to impact a wide range of applications in medical image analysis and beyond. There is still much to be explored and improved in this area, and we hope that our work will inspire future research in open-set source-free domain adaptation.

5.2 Future Directions

The topic of open-set source-free domain adaptation in fundus image analysis is a highly relevant and interesting area of research, and there is much potential for further work in this field. Some possible directions for future research in this area include:

- Improved domain adaptation techniques: There are many existing techniques for domain adaptation, but they may not always be effective in the open-set source-free scenario. There is a need for new methods that can effectively adapt to novel domains without access to source data.

- Extension to other medical imaging modalities: The techniques developed in this work can potentially be extended to other medical imaging modalities such as CT and MRI. This would allow the use of these methods in a wider range of medical applications.

- Incorporation of additional data sources: While this work focused on adapting to a new domain using only the target data, it may be possible to incorporate other data sources (such as unannotated data or auxiliary tasks) to improve performance.

- Evaluation on larger and more diverse datasets: The datasets used in this work were relatively small and may not fully represent the variability in real-world fundus images. Future work should evaluate these techniques on larger and more diverse datasets to better understand their generalizability.

- Integration with other tasks: Domain adaptation techniques can be used in combination with other tasks such as segmentation or diagnosis. Future work could explore the integration of these methods with other tasks to improve the overall performance of the system.

In general, there is much potential for further research in the area of open-set source-free domain adaptation in fundus image analysis. These directions represent only a few of the many possibilities for future work, and there is much room for innovation and exploration in this field.

Bibliography

[1] Fighting blindness canada website. `https://www.fightingblindness.ca/`.

[2] International competition on ocular disease intelligent recognition (2019). `https://odir2019.grand-challenge.org`.

[3] Tencent miying artificial intelligence competition for medical imaging (2021). `https://contest.taop.qq.com/`.

[4] Who launches first world report on vision. `https://www.who.int/blindness/Vision2020_report.pdf`.

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.

[6] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Silvia Bucci, Mohammad Reza Loghmani, and Tatiana Tommasi. On the effectiveness of image rotation for open set domain adaptation. In *European Conference on Computer Vision*, pages 422–438. Springer, 2020.

[8] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019.

[9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 1251–1258, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[13] S Gayathri, Varun P Gopi, and Ponnusamy Palanisamy. A lightweight cnn for diabetic retinopathy classification from fundus images. *Biomedical Signal Processing and Control*, 62:102115, 2020.

[14] Yosief Gebremariam. Detection of diabetic retinopathy using deep convolutional neural network on mobile devices, 2022.

[15] Luca Giancardo. *Automated fundus images analysis techniques to screen retinal diseases in diabetic patients*. PhD thesis, Université de Bourgogne, 2011.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

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

[18] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998. Pmlr, 2018.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[23] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020.

[24] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12376–12385, 2020.

[25] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9757–9766, 2021.

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[27] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] Jinke Lin, Qingling Cai, and Manying Lin. Multi-label classification of fundus images with graph convolutional network and self-supervised learning. *IEEE Signal Processing Letters*, 28:454–458, 2021.

[29] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019.

[30] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[31] Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. *Data*, 6(2):14, 2021.

[32] Yingwei Pan, Ting Yao, Yehao Li, Chong-Wah Ngo, and Tao Mei. Exploring category-agnostic clusters for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13867–13875, 2020.

[33] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-second AAAI Conference on Artificial Intelligence*, 2018.

[34] Sehrish Qummar, Fiaz Gul Khan, Sajid Shah, Ahmad Khan, Shahaboddin Shamshirband, Zia Ur Rehman, Iftikhar Ahmed Khan, and Waqas Jadoon. A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, 7:150530–150539, 2019.

[35] Joshna S and P Student. International journal of innovative research in computer and communication engineering a study on diabetic retinopathy using exudate segmentation. *International Journal of Innovative Research in Computer and Communication Engineering*, 6:1119–1125, 02 2018.

[36] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR, 2017.

[37] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018.

[38] Hiram José Sandoval-Cuellar, Gendry Alfonso-Francia, MA Vázquez-Membrillo, Juan Manuel Ramos-Arreguín, and S Tovar-Arriaga. Image-based glaucoma classification using fundus images and deep learning. *Revista mexicana de ingeniería biomédica*, 42(3), 2021.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer vision*, pages 618–626, 2017.

[40] Tasfia Shermin, Guojun Lu, Shyh Wei Teng, Manzur Murshed, and Ferdous Sohel. Adversarial network with multiple classifiers for open set domain adaptation. *IEEE Transactions on Multimedia*, 23:2732–2744, 2020.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[43] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[44] Jianqing Wang, Weitao Mo, Yan Wu, Xiaomei Xu, Yi Li, Jianming Ye, and Xiaobo Lai. Combined channel attention and spatial attention module network for chinese herbal slices automated recognition. *Frontiers in Neuroscience*, 16, 2022.

[45] Jing Wang, Liu Yang, Zhanqiang Huo, Weifeng He, and Junwei Luo. Multi-label classification of fundus images with efficientnet. *IEEE Access*, 8:212499–212508, 2020.

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[48] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[49] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019.

[50] Shuang Yu, Kai Ma, Qi Bi, Cheng Bian, Munan Ning, Nanjun He, Yuexiang Li, Hanruo Liu, and Yefeng Zheng. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 45–54. Springer, 2021.

[51] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[52] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2736–2746, 2022.

[53] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[54] Youshan Zhang. A survey of unsupervised domain adaptation for visual recognition. *arXiv preprint arXiv:2112.06745*, 2021.

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[56] Yi Zhou, Shaochen Bai, Tao Zhou, Yu Zhang, and Huazhu Fu. Delving into local features for open-set domain adaptation in fundus image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 682–692. Springer, 2022.

Curriculum Vitae

Name:	Masoud Pourreza
Post-Secondary Education and Degrees:	BSc. in Computer Engineering 2011 - 2015 Semnan University Semnan, Iran
Related Work Experience:	Graduate Teaching and Research Assistant The University of Western Ontario 2021 - 2022

Publications:

1. **Pourreza, M.**, Mohammadi, B., Khaki, M., Bouindour, S., Snoussi, H., Sabokrou, M. (2021). G2D: generate to detect anomaly. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 2003-2012).

2. Sabokrou, M., **Pourreza, M.**, Li, X., Fathy, M., Zhao, G. (2021). Deep-HR: Fast heart rate estimation from face video under realistic conditions. Expert Systems with Applications, 186, 115596.

3. Sabokrou, M., **Pourreza, M.**, Fayyaz, M., Entezari, R., Fathy, M., Gall, J., Adeli, E. (2018, December). AVID: Adversarial visual irregularity detection. In Asian Conference on Computer Vision (pp. 488-505). Springer, Cham.

4. **Pourreza, M.**, Salehi, M., Sabokrou, M. (2021). Ano-graph: Learning normal scene contextual graphs to detect video anomalies. arXiv preprint arXiv:2103.10502.

5. **Pourreza, M.**, Derakhshan, R., Fayyazi, H., Sabokrou, M. (2018, December). Sub-word based Persian OCR using auto-encoder features and cascade classifier. In 2018 9th International Symposium on Telecommunications (IST) (pp. 481-485). IEEE.