

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Unsupervised Domain Adaptation in Medical Images

Author:
Wenqing Zong

Supervisor:
Dr. Matthew Williams
Dr. Elsa Angelini

Submitted in partial fulfillment of the requirements for the MSc degree in
Advanced Computing of Imperial College London

October 2022

Abstract

Deep learning has made remarkable achievements in the field of medical image analysis, but most research results must be supervised training using labelled datasets, but labels of medical images are often difficult to obtain, which makes it difficult to translate research results into practical applications. At the same time, the domain shift problem makes it difficult for a model trained on one dataset/modality/institution to maintain similar performance on another dataset/modality/institution. This project attempts to use the Unsupervised Domain Adaptation (UDA) method to solve the above two difficulties. In terms of the specific problem, this project focuses on the two most common tasks in medical image analysis: classification and segmentation.

In the Classification task, this project uses the CrossModa 2022 dataset and the Cross Domain Transformer (CDTrans) network structure. CDTrans was proposed by Xu et al. in March of this year and is one of the earliest works to solve UDA problem via transformer idea. However, the results show that it is unsuitable for the CrossModa 2022 dataset. In the Segmentation task, this project uses the BraTS 2021 dataset. Inspired by Chen et al., this project proposes a novel Momentum Prototype UDA (MP-UDA) procedure and achieves better results when compared with the latest UDA work, Black Box UDA. At the same time, MP-UDA is designed with data privacy in mind. In cross-institution cooperation, MP-UDA can completely avoid data sharing, which requires complex paperwork and long approval processes.

Acknowledgments

Time flies. Completing this thesis marks the end of my Master's studies at Imperial. Although it's only one year, I'm still deeply impressed by Imperial College London. Here I met knowledgeable professors, enthusiastic classmates, and gained precious friendships. I love everything here.

For this thesis, first of all, I would like to thank my supervisor Dr. Matthew Williams, who is always very patient in answering my doubts and offering me guidance. I also would like to thank my second marker Dr. Elsa Angelini, who gave me invaluable advice during the initial stage of the project when I was struggling to find a suitable dataset.

Secondly, I would like to thank my parents, my girlfriend and my lovely pet rabbit Sleepy. When I was working on this project, I often feel incomparable pressure, and it was you who provided me with spiritual comfort.

Contents

| | |
|--|-------------|
| List of Figures | vii |
| List of Tables | viii |
| 1 Introduction | 1 |
| 1.1 Motivation | 1 |
| 1.1.1 Medical Images | 1 |
| 1.1.2 Medical Images with Traditional Machine Learning | 2 |
| 1.2 Goal of This Project | 5 |
| 1.3 Report Overview | 5 |
| 2 Literature Review | 7 |
| 2.1 Definition of Unsupervised Domain Adaptation | 7 |
| 2.2 Taxonomy of UDA | 8 |
| 2.3 Recent Approaches to UDA Problem | 9 |
| 2.3.1 Four Common Approaches in Transfer Learning | 9 |
| 2.3.2 Recent UDA Approaches | 10 |
| 3 Material and Evaluation Metrics | 13 |
| 3.1 Data | 13 |
| 3.1.1 CrossModa 2022 Challenge Data | 14 |
| 3.1.2 BraTS 2021 Data | 16 |
| 3.2 Evaluation Metrics | 17 |
| 3.2.1 Evaluation Metrics for Classification | 17 |
| 3.2.2 Evaluation Metrics for Segmentation | 18 |
| 4 Methodologies | 19 |
| 4.1 Classification | 19 |
| 4.1.1 Network Architecture | 20 |
| 4.1.2 Training Procedure | 21 |
| 4.2 Segmentation | 22 |
| 5 Experiment Results | 25 |
| 5.1 Hardware and Software Environment Setup | 25 |
| 5.2 Classification Result | 25 |

| | |
|---|-----------|
| 5.3 Segmentation Result | 30 |
| 6 Conclusions and Future Work | 37 |
| 6.1 Conclusions | 37 |
| 6.2 Future Work | 38 |
| 7 Ethical Considerations | 39 |
| 7.1 Data Ethics | 39 |
| 7.1.1 CrossModa 2022 Dataset Privacy Protection | 39 |
| 7.1.2 BraTS 2021 Dataset Privacy Protection | 40 |
| 7.2 Method Ethics | 41 |
| 7.3 Code Ethics | 41 |
| Bibliography | 42 |
| A Imperial Ethics Checklist | 47 |

List of Figures

| | | |
|------|---|----|
| 1.1 | Four image tasks that can neural networks can solve: Image Classification, Object Detection, Semantic Segmentation, Instance Segmentation | 2 |
| 1.2 | Four Classes in Koos Classification System | 3 |
| 1.3 | Brief Schematic of SRS Treatment | 4 |
| 3.1 | CrossModa 2022 Dataset Examples | 14 |
| 4.1 | Network Architecture of CDTrans | 20 |
| 4.2 | Data Flow in MP-UDA. In this figure, black boxes are data, blue boxes are operations and green boxes are model parameters | 23 |
| 5.1 | CDTrans Pre-train Loss on CrossModa 2022 Source Domain Training Set | 27 |
| 5.2 | Three Classification Metrics of Koos Classification of CDTrans on CrossModa 2022 Source Domain Training Set | 27 |
| 5.3 | Precision of Koos classification of CDTrans on CrossModa 2022 Source Domain Validation Set | 28 |
| 5.4 | CDTrans Loss On CrossModa 2022 Target Domain Training Set | 29 |
| 5.5 | CrossModa 2022 Official Leader-Board for Classification Task | 30 |
| 5.6 | MP-UDA Dice Coefficient During BraTS 2021 Source Domain Training | 32 |
| 5.7 | MP-UDA Performance on BraTS 2021 Validation Set | 32 |
| 5.8 | MP-UDA Loss On BraTS 2021 Target Domain Training Set | 33 |
| 5.9 | MP-UDA True Dice Coefficient On BraTS 2021 Target Domain Training Set | 34 |
| 5.10 | Visualisation of MP-UDA Performance | 36 |
| 7.1 | CrossModa 2022 London Data Blurs Patients' Facial Information | 40 |
| 7.2 | CrossModa 2022 Tilburg Data Crops Patients' Facial Information | 40 |
| 7.3 | BraTS 2021 Data is Skull Stripped | 41 |
| A.1 | Imperial Ethical Checklist Part 1 | 48 |
| A.2 | Imperial Ethical Checklist Part 2 | 49 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | Class Distribution of The CrossModa 2022 Source Domain Data . . . | 15 |
| 3.2 | Meta Data on Two Medical Institutions | 16 |
| 5.1 | CDTrans Hyper-parameters | 26 |
| 5.2 | MP-UDA Hyper-parameters and Data Augmentation | 31 |
| 5.3 | Comparison With BBUDA and BraTS 2021 SOTA | 35 |
| 5.4 | Modifications Made To The Re-implemented BBUDA | 35 |

Chapter 1

Introduction

Section 1.1.1 introduces some basic knowledge of medical images. Then a brief introduction to the type of image tasks that machine learning can accomplish will be given in section 1.1.2, it also contains a brief analysis of why traditional machine learning methods have little effect on medical images. The above two sub-sections together form the motivation of this project. Section 1.2 outlines the aim of this project from an application point of view. Moreover, section 1.3 describes how the other chapters are organised in this thesis.

1.1 Motivation

1.1.1 Medical Images

According to [1], medical images are produced by irradiating the human body with some form of energy, which is absorbed or scattered by the internal tissues of the human body, and then received by external instruments.

Depending on the imaging physics principle, medical images can be subdivided into different types, and we use the word “modality” to refer to these different kinds. Some common modalities are CT images (obtained via body tomography by X-ray), and MR images (obtained by the principle of nuclear magnetic resonance). Among them, by changing some specific nuclear magnetic resonance parameters, MR images can be further divided into T1 weighted, T2 weighted, and FLAIR images [1]. One special MRI modality is T1 contrast enhanced (T1ce), which requires an injection of Gadolinium to reveal new lesions. However, as described by Cancer Research UK [2], not everyone is suitable to take a T1ce scan as they may be allergic to the contrast medium.

At the same time, it should also be noted that the imaging physics principle also affects the difficulty of acquiring images in different modalities. For example, MRI images require a longer scanning time and are more expensive when compared with CT images.

As summarised by [3], different modalities have different suitable usage scenarios. For example, T1 is sensitive to new lesions (i.e., active areas of disease), and T2 is used to show old, inactive lesions.

1.1.2 Medical Images with Traditional Machine Learning

In recent years, neural networks have attracted the interest of many researchers and have made remarkable achievements in the field of image analysis. Figure 1.1 shows the 4 different kinds of image processing tasks that neural networks can solve. The four tasks are: (in order of difficulty from easy to hard):

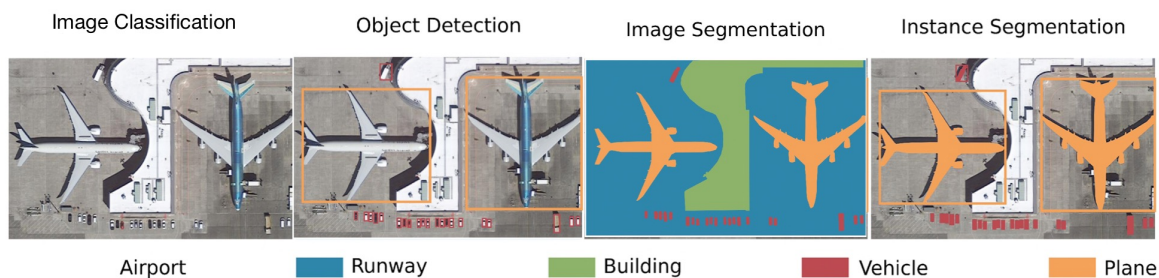


Figure 1.1: Four image tasks that can neural networks can solve: Image Classification, Object Detection, Semantic Segmentation, Instance Segmentation

- **Image Classification:** Predicts which object class is presented in the input image.
- **Object Detection:** Predicts which object class is presented in the input image and also predicts a bounding box to show its general location.
- **Semantic Segmentation:** This is a bit like a pixel-wise classification problem, where we assign a class label to each pixel in the input image. If the input image contains two or more instances of the same object class, then different instances receive identical label.
- **Instance Segmentation:** This is an extension of Semantic Segmentation. Each instance of the same object class receives different labels (shown as adding an extra bounding box in Fig 1.1).

Among them, classification and semantic segmentation have extensive applications in the medical image analysis. Two example applications are given for each task below:

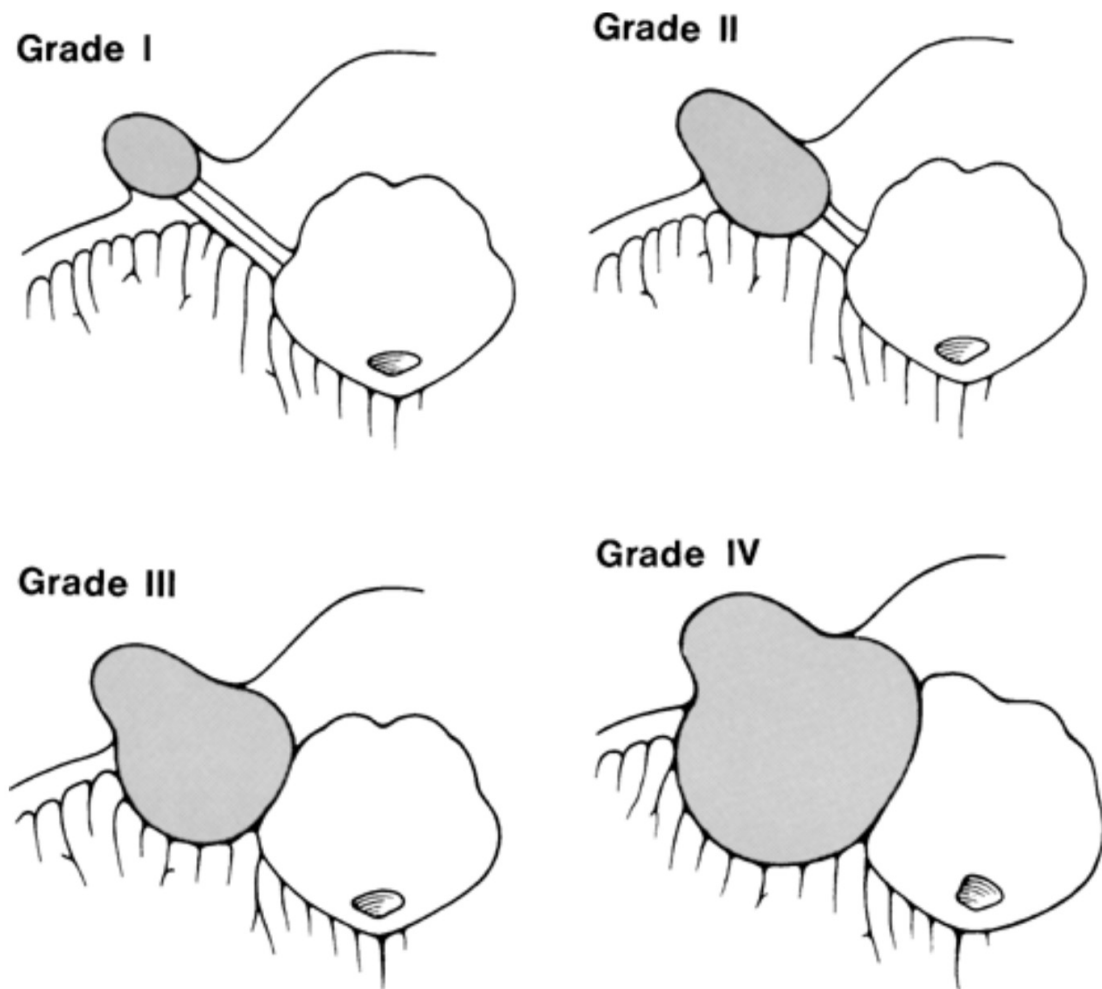


Figure 1.2: The four classes in Koos classification system according to [4], where the gray circle represents VS and the white circle represents brainstem

- For classification task. According to NHS, Vestibular Schwannoma (VS) is ‘a type of non-cancerous (benign) brain tumour’ [5]. It usually starts from the hearing nerve and grows slowly over time. In 1998, Koos [4] proposed a classification system to determine the severeness of VS and concluded that the tumor size, along with patients’ preoperative hearing ability, have a huge impact on hearing preservation after surgery. The Koos classification system divides the severity of VS into four categories based on factors such as tumor size and whether it touches/compresses the brainstem. Figure 1.2 shows an ideal representation of the four classes in the Koos classification system (taken from the original paper). In recent years, the Koos classification system is also used for treatment planning [6].

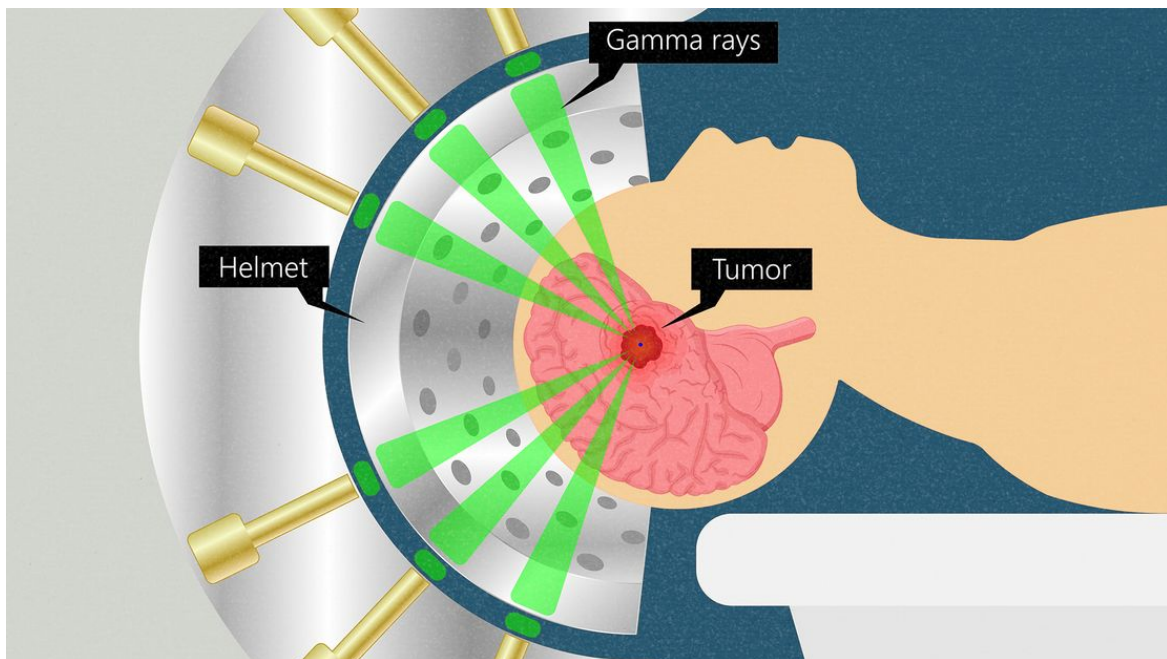


Figure 1.3: An intuitive schematic of SRS treatment taken from Roan’s article [7], lesion is labeled in red and healthy brain is labeled in pink

- For segmentation task. A simple treatment for some tumour diseases, such as brain metastasis, is Whole-Brain Radiation (WBR). This method uses a high-energy beam to irradiate the entire brain in the hope of killing tumour cells. One drawback of this method is that it indiscriminately attacks all cells in the skull, as a result, it sometimes injures healthy cells accidentally. An improvement on this method is Stereotactic Radios Surgery (SRS), which uses many small doses of radiation, but delivers large doses of radiation at the point where the rays meet to kill tumour cells [8]. An increasing number of doctors are using SRS in recent years, and, according to Shinde et al. [9], SRS is believed to achieve better results than WBR. Figure 1.3 provides an intuitive illustration of how SRS works, it also reveals the fact that SRS is actually a segmentation task as the doctors need to distinguish whether a voxel is normal brain tissue to be preserved or tumor cells that should be killed.

Ideally, both of the above examples can be solved using traditional deep learning methods and are expected to achieve ideal performance, if and only if we have sufficient training data and corresponding labels. However, it is not always the case in reality. In the real world, when researchers want to turn their academic fruits into actual commercial products, they might encounter three problems:

1. Lack of labels. It is difficult to obtain medical image labels. For general RGB image segmentation tasks, ordinary people can accurately manually label each image without any training. However, in the field of medical images, ordinary people are completely incompetent, and the data can only rely on relevant expert experience. Also, for segmentation tasks, manually labelling can be time-consuming as most medical images are 3D.

2. Domain shift decreases model performance. Medical data are susceptible to the parameters of the equipment. Even if two institutions use equipment from the same manufacturer, and adopt the same modality, the resulting data distribution may still differ due to subtle parameter differences during operation. As a result, model trained in one institution cannot directly be used in another. Also, compared to cross-institution scenario, domain shift is more severe in cross-modality situation.

Even within the same institution, model trained on one modality will not work well on another modality. However, it's hard for some people to get multi-modality data, one example is patient being allergic to Gadolinium so their T1ce image is unavailable.

3. Privacy makes data sharing difficult. Although the above domain shift problem could be alleviated by providing the model with data from two domains simultaneously, but data sharing could make this option invalid. All medical data, especially image data, should be considered as patient privacy. As a result, medical data sharing usually requires an anonymisation preprocess, complex paperwork and lengthy approval process.

1.2 Goal of This Project

This project aims to solve the problem of classification and segmentation of medical images under the influence of the above mentioned three unfavourable factors. Specifically, we establish the following restrictions:

1. Data come from two different domains; only one domain contains the corresponding ground truth labels. We will call the domain which has labels as 'Source Domain', and the other as 'Target Domain'.
2. There is an obvious domain shift between two domains so a model trained with only one domain data cannot be directly applied to the other.

1.3 Report Overview

This chapter mainly introduces the basic knowledge of medical images, reviews the image problems that traditional machine learning methods can solve, and briefly analyses three factors that affect machine learning model performance in the practical application of medical images. This chapter also formally presents the problem that this project is expected to solve.

Chapter 2 will first formally define the Unsupervised Domain Adaptation (UDA) problem from a mathematical point of view and subdivide the UDA problem according to different constraints. As UDA is a subfield of Transfer Learning (TL) we will review four common ideas to solve TL and then present some recent works specially aimed at UDA.

In the third chapter, we will introduce the data used in this project, including data sources, basic information, and how the dataset is split. At the same time, chapter 3 also introduces the evaluation metrics that will be used in this project.

The fourth chapter will detail the specific methods used by this project to solve the classification and segmentation UDA problems, including the overall network architecture and training process. It should be noted that the solution to the segmentation UDA problem is newly proposed by this project and is one of the leading academic contributions.

Chapter 5 will first introduce the project's hardware configuration and software environment and then analysis the results of each step in the training process and discuss what it means.

Chapter 6 will review the success and failure of the entire project and look forward to future research directions.

Finally, Chapter 7 will review three ethical problems we met during this project and how they are tackled.

Chapter 2

Literature Review

This chapter will first give the mathematical definition of Domain Adaptation, introduce a taxonomy in detail. As Unsupervised Domain Adaptation is a sub-field of Transfer Learning (TL), 4 common approaches

2.1 Definition of Unsupervised Domain Adaptation

In the Unsupervised Domain Adaptation (UDA) task, we aim to transfer knowledge to an unlabeled target domain after learning from a labelled source domain.

Various mathematical definition exist for UDA in the existing literature, but some seem to use a simplified version designed for specific situations. For example, Wang’s [10] definition of UDA does not support multiple source domains, and Csurka’s [11] definition does not support open-set setting. A detailed Taxonomy of UDA will be provided in section 2.2. Here, we will adopt the definition of Zhao et al. [12] as this version is the most general one.

We’ll start with defining a set of all inputs:

$$\mathcal{X} = \{x_i\}_{i=1}^N \quad (2.1)$$

where x_i is one input image and N is the total number of images in this set.

Then, the label set is defined as:

$$\mathcal{Y} = \{y_i\}_{i=1}^N \quad (2.2)$$

Depending on different tasks, y_i may be drawn from \mathcal{R} for classification problem, or it could be drawn from \mathcal{R}^{H*W*C} for segmentation problem, and H, W represents the output image spatial dimension, whereas C represents the total number of classes.

A domain is the combination of paired input samples their corresponding labels, drawn from a distribution p :

$$\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}, \text{ and they follow a distribution } p(\mathcal{X}, \mathcal{Y}) \quad (2.3)$$

Now, UDA problem can be formally defined as: given n_s source domains \mathcal{S}_1 to \mathcal{S}_{n_s} , the model should learn on these input-label pairs and adapt the learnt knowledge to another target domain \mathcal{T} , where the target domain label is not available for the model to learn.

2.2 Taxonomy of UDA

Based on the above definition, UDA can be further split into different categories. Some possible further taxonomy includes:

Divide by number of source domains.

- Single Source UDA: where $n_s = 1$,
- Multi Source UDA: where $n_s > 1$.

Divide by data dimension. Suppose $x_i^k \in \mathcal{R}^{d_i}$ represents the k th sample from the i th source domain has spatial dimension d_i , and $x_{\mathcal{T}}^k \in \mathcal{R}^{d_{\mathcal{T}}}$ means the k th sample from the target domain has spacial dimension $d_{\mathcal{T}}$. Then:

- Homogeneous UDA: where $d_i = d_j = d_{\mathcal{T}}$
- Heterogeneous UDA: otherwise.

Divide by label space: suppose \mathcal{C}_i and $\mathcal{C}_{\mathcal{T}}$ represents the label space of the i th source domain and the target domain, respectively, then:

- Closed-Set UDA: where $\mathcal{C}_i = \mathcal{C}_j = \mathcal{C}_{\mathcal{T}}$.
- Open-Set UDA: where at least one source domain contains some target label classes: $\exists \mathcal{C}_i (\mathcal{C}_i \cap \mathcal{C}_{\mathcal{T}} \subset \mathcal{C}_{\mathcal{T}})$.
- Partial UDA: where at least one source domain contains all target label classes: $\exists \mathcal{C}_i (\mathcal{C}_{\mathcal{T}} \subset \mathcal{C}_i)$

Divide by the number of target domains:

- Multi Target UDA: where there are more than 1 target domain.

The above taxonomy divides UDA problem from a mathematical point of view. In 2020, Liang et al. [13] focus on UDA training process and found that after the neural network has learned on the labelled source domain data, it is unnecessary to re-access the source domain data while migrating learnt knowledge to the target domain. They defined this approach as Source Free UDA. Subsequently, researchers [14, 15] found that not only the source domain data could be hidden for target domain training, but the source domain model can also be decoupled from initialising the target domain model. This approach is defined as Black Box UDA.

In summary, a recent trend is to divide UDA by accessibility of source data/model:

- Source Free UDA: The source data is only used to train the source model and not accessible while training the target model. This is also called White Box

UDA as the source model is still available.

- **Black Box UDA:** After training on the source domain, both source data and source model are hidden for target domain training. Only an API is provided to access the source model input and output.

It is worth noting that in Black Box UDA setting, the source model and target model could use different network architecture, so it provides more flexibility.

2.3 Recent Approaches to UDA Problem

2.3.1 Four Common Approaches in Transfer Learning

UDA is actually a sub-field of Transfer Learning (TL), and we noticed that TL methods still guide recent UDA research to some extent, so it is necessary to focus on 4 common TL methods first.

Instance Based

The core idea of this kind of method lies in the fact that there must be some source domain data looks similar to the target domain. We can maximise the similarity between two domains by assigning more weight to those source domain samples, thus transferring knowledge to the target domain. Khan et al. [16] propose a new iterative method to learn a common subspace to do cross-domain learning by using non-parametric quadratic mutual information. Tan et al. [17] put forward a selective learning algorithm. This algorithm requires a middle domain as the bridge to transfer knowledge to the target domain, it selects unlabeled middle domain data which has a positive effect on target domain learning.

Feature Based

This type of method narrows the feature distribution of two domains by doing feature transformations. Pan et al. [18] find a novel dimensionality reduction method that maps data from two domains into a regenerated kernel Hilbert space. In this space, the source and target data distance is minimised while their respective internal properties are preserved. Duan et al. [19] propose that cross-domain learning could be done by using multiple kernel functions. Their method learns the kernel function by minimizing the structural risk function and the distribution mismatch between domains. Long et al. [20] also follow the idea of dimensional reduction and combine feature based method and instance re-weighting. Their method, Transfer Joint Matching (TJM), matches features at low dimensions while re-weighting each data sample to reduce domain shift. Their constructed new feature representations are insensitive to distribution differences and irrelevant instances. Zhang et al. [21] minimize domain shift by reducing the statistical and geometric differences between the features of two domains to transfer knowledge.

Model Based

This method is the most common one in the deep learning era. The core idea is that the model parameters represent the learned knowledge, and sharing the parameters can achieve knowledge transfer between two domains. Yonsinski et al. [22] discuss the transferability of deep neural networks and conclude that: for a deep network, as the layer deepens, the network becomes more and more task dependent while shallow layers only learn general features. Tajbakhsh et al. [23] explore the possibility of fine-tuning a pre-trained CNN network to fit medical datasets and answer the question of which layers to fine-tune. They also proved that fine-tuning could achieve better performance than training from the beginning. Ghifary et al. [24] propose Domain Adaptive Neural Network (DANN), which adds an Maximum Mean Discrepancy (MMD) adaptation layer after the feature layer to minimize the distance between source and target domain. A similar idea is also adopted by Tzeng et al. [25] in their work Deep Domain Confusion (DDC). However, Tzeng et al. also use an additional domain confusion loss to learn domain invariant feature representations. Compared to the DDC, Long et al. [26] added two more modifications to their work Deep Adaptation Network (DAN): they use more adaptation layers rather than one, and they replace MMD with multi-kernel MMD. Long et al. [27] also further optimize their own work later by considering the joint distribution of feature and label.

Relation Based

This type of method is mainly used to learn common (logical) relations between source and target domain. It receives less attention in the deep learning era. To our best knowledge, Pan et al. [28]’s survey is the last one that contains some methods for relational transfer learning. It introduces Mihalkova et al. [29]’s and Davis et al. [30]’s work, where the former built a Markov Logic Networks to find similarities among different domains and the latter used second-order Markov logic to transfer relational knowledge. Later, in a 2016 survey [31], the author mentioned Li et al. [32]’s relation-based work on text classification, but this method does not generalise well to other non-text tasks. In a 2019 survey [33], relation-based transfer learning is completely omitted.

2.3.2 Recent UDA Approaches

In addition to the above four common categories, several new ideas have emerged for UDA tasks in recent years. Specifically, they can be divided into adversarial-based, source-free based and frequency-based. Strictly speaking, adversarial-based and source-free-based methods still belong to the model-based method, but we think that listing them separately will help readers understand the latest direction in this field. With the popularity of deep learning in recent years, it can be considered that almost all method needs to train a model. Therefore, the general ‘Model-Based’ term cannot precisely reflect researchers’ novelty.

Adversarial Based

Ganin et al. [34] are one of the first researchers to use adversarial ideas to solve UDA tasks. They proposed a novel network architecture: domain-adversarial neural network (DANN). The main idea of this network is somewhat similar to GAN. The DANN network consists of a feature extractor, a label predictor, and a domain classifier. The purpose of the feature extractor is to extract features from the input image and send them to the label predictor network to form a complete feed-forward process. At the same time, the extracted features will also be sent to the domain classifier, whose task is a simple binary classification problem to determine which domain the input features come from. During training, the loss of the label predictor is minimised to ensure the model's performance on the source domain, and the loss of the domain classifier is also minimised to encourage the network to extract domain invariant features. The gradient information from the domain classifier is passed to the feature extractor after a gradient reversal layer.

Due to the excellent performance of CycleGan on image style transfer, researchers have tried to use it to solve the UDA problem. Specifically, CycleGan is used to convert the source domain images into target domain images and then use these labelled generated target domain images to train the segmentation network. However, it should be noted that the images generated by CycleGan only have target domain style and often lose source domain details. In response to this shortcoming, Jiang et al. [35] propose a new loss function, including tumour loss and feature loss. The tumour loss tries to make the model achieve similar segmentation results on CT (its source domain) and generated MRI images (its target domain). The feature loss tries to make the two segmentation networks share the same high-level feature to ensure that the model does not miss small tumours. Cai et al. [36] also pay attention to training loss design. In addition to adversarial loss and cycle consistency loss, they added shape-consistency loss to ensure that the generator gives shape-invariant target domain images. Sankaranarayana et al. [37] focuses on the input of the GAN generator (G). They use the feature extractor (F) to extract features from images, and use its output as the generator (G) input. G aims to generate fake images in the same domain, while discriminator (D) distinguishes real and fake images. It is worth noting that the training goals of the three networks G, D, and F are different: the goal of G is to make the fake images look like real images, the goal of D is to discriminate between fake and real images, and the goal of F is to provide image embedding such that the fake image of G is considered to be real image of the other domain. In other words, G and D focus on the same domain, while the goal of F is on cross-domain. As G's output becomes more and more challenging to distinguish for D, F gradually extracts domain invariant feature representation.

Source-Free Based

Source-free UDA has been a new research topic in the past two years. In this UDA setting, the source domain data is only used to train the source domain model and will not be re-accessed in subsequent UDA training. This idea is first demonstrated possible by Liang et al. [13] In their method, the source domain model consists of

two parts: a feature extractor and a classifier. The source feature extractor weight is used to initialise the target feature extractor, but the classifier is locked, so its weight cannot be modified during target domain training. Subsequently, based on this UDA setting, Chen et al. [38] further explore how to denoise pseudo-labels. Their denoising method consists two parts: 1. Uncertainty Map. Monte Carlo Dropout is adopted in source model inference. The same image is inferred multiple times to obtain different results. Pixel-wise variance is calculated and treated as uncertainty. The larger the variance, the more uncertain the pixel is. 2. Prototype Estimation, the author regards the input of the last convolutional layer as ‘feature’ and the output of the last convolutional layer as ‘probability’. Prototype features can be obtained by multiplying these two items. After that, the denoised pseudo label is obtained by a pixel-wise distance measure. Liu et al. [39] noticed that the Batch Normalisation (BN) layer contains domain-specific information and uses exponential momentum decay to gradually update BN mean and variance to adapt to the target domain. In a subsequent paper, Zhang et al. [15] propose black box UDA scenario where an additional limitation is added to the source-free UDA setting: source domain model weight cannot be obtained, and only the input and output APIs are provided instead. Liu et al. [14] give a promising solution to black box UDA problem: they use the weighted sum of source model and target model output to construct pseudo-labels, and exponentially decrease the contribution of the source domain model to achieve knowledge distillation.

Source-free UDA is a new research direction in the past two years. Compared with other UDA methods, source-free UDA provides greater flexibility, making it more likely to be adopted in real-world practical applications. That is the reason why this project chooses to build a novel source-free UDA method. The new Momentum Prototype UDA (MP-UDA) is based on the work of [38] and [13]. Details of this method will be introduced in Section 4.2.

Frequency Based

As opposed to understanding images from pixels, another possible idea is to understand images from frequency. Fourier transformation can extract features from the frequency perspective: high frequency extracts small details of the image, and low frequency extracts the overall image structures. Yang et al. [40] adopt this idea and propose the Fourier Domain Adaptation (FDA) architecture. This method aligns domain distribution by swapping the low-frequency spectrum between the source and target domains. The FDA method does not require any training to learn domain invariant features, it only requires Fourier transform and its inverse.

Chapter 3

Material and Evaluation Metrics

This chapter will give a detailed description of the data used in this project and introduce the evaluation metrics.

3.1 Data

Since this project focuses on UDA applications for both classification and segmentation tasks, two different datasets are used: CrossModa2022 [6] for classification problems and BraTS2021 [41, 42, 43] for segmentation problems.

3.1.1 CrossModa 2022 Challenge Data

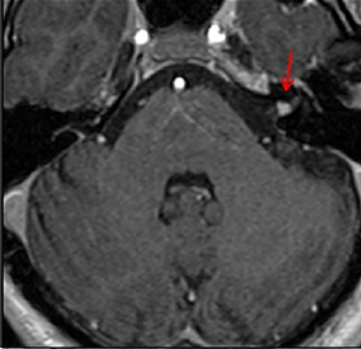
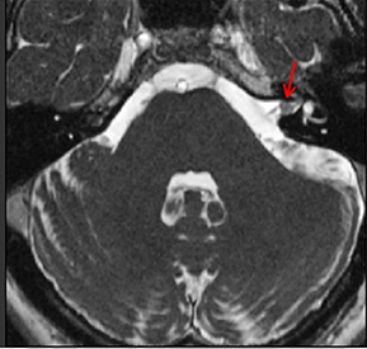
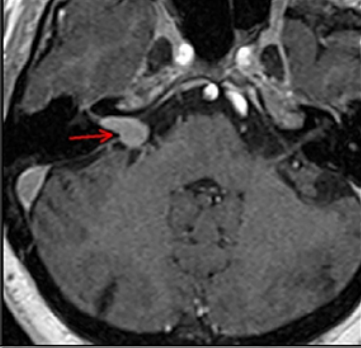
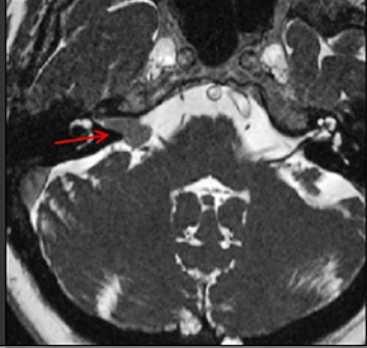
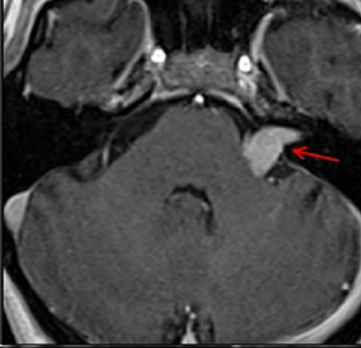

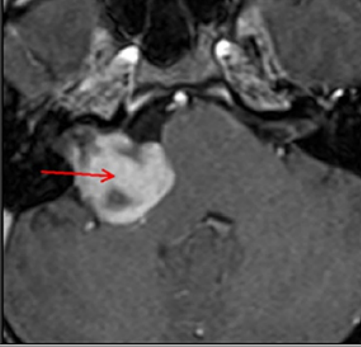
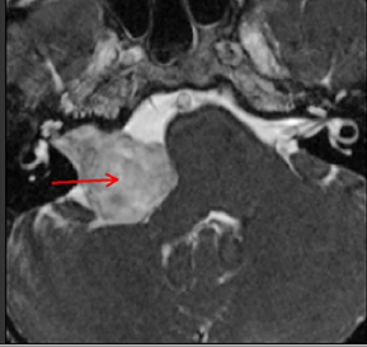
| Grade | Criteria | Representative ceT1 image | Representative hrT2 image |
|------------|--|---|---|
| I | Tumours are completely confined to the internal auditory canal. |  |  |
| II | Tumours have both intra- and extrameatal components, extending into the cerebellopontine angle (CPA) but do not contact the brainstem. |  |  |
| III | Tumours contact the brainstem but do not compress it. |  |  |
| IV | Tumours cause brainstem compression and/or displacement of adjacent cranial nerves. |  |  |

Figure 3.1: Example Source Domain (T1ce) and Target Domain (hrT2) images for each Koos grades [44]

The problem is to determine each subject's Vestibular Schwannoma (VS) grade according to the Koos grading system [4]. This dataset contains contrast-enhanced

T1 MRI and high-resolution T2 MRI images from two different medical institutions, which are referred to as London data and Tilburg Data. Among them, the contrast-enhanced T1 MRI is defined by the challenge organiser as the source domain, which contains the corresponding label, and the high-resolution T2 MRI is the Target domain, which has no corresponding label. The challenge organizers [44] provides some example images and their corresponding labels in Figure 3.1. There are altogether 210 training samples in the source domain, but 28 of them are post-operative data so they are eliminated for the classification task, resulting only 182 source domain training samples. The class distribution can be found in Table 3.1:

Table 3.1: Class Distribution of The CrossModa 2022 Source Domain Data

| | | | |
|---------|----------|-----------|----------|
| Grade I | Grade II | Grade III | Grade IV |
| 12 | 51 | 73 | 46 |

On the target domain side, CrossModa 2022 dataset contains 210 training samples, but no label information is provided, and the whole dataset is organised in an unpaired way, so no subject would appear both in the source and the target domain.

CrossModa 2022 also provides 64 validation samples for participants, but the validation set label is not available. Instead, the challenge provides Macro Average-Mean Absolute Error (MA-MAE) score as (the only) feedback. MA-MAE will be formally defined in section 3.2.1.

CrossModa 2022 also detailed some medical parameters they used to acquire the data, and they are summarised as follows in Table 3.2:

Table 3.2: Meta Data on Two Medical Institutions

| | London Data | Tilburg Data |
|----------------------|---|---|
| Manufacture | 32-channel Siemens Avanto 1.5T scanner using a Siemens single-channel head coil | Philips Ingenia 1.5T scanner using a Philips quadrature head coil |
| T1ce sequence | MPRAGE | 3D-FFE |
| T1ce resolution | 0.4*0.4mm, 512*512 | 0.8*0.8mm, 256*256 |
| T1ce slice thickness | 1.0 - 1.5mm | 1.5mm |
| T1ce TR | 1900ms | 25ms |
| T1ce TE | 2.97ms | 1.82ms |
| T1ce TI | 1100ms | N/A |
| T2 Sequence | 3D CISS or FIESTA | 3D-TSE |
| T2 resolution | 0.5*0.5mm, 384*384 or 448*448 | 0.4*0.4mm, 512*512 |
| T2 slice thickness | 1.0 - 1.5mm | 1.0mm |
| T2 TR | 9.4ms | 2700ms |
| T2 TE | 4.23ms | 160ms |
| T2 ETL | N/A | 50 |

3.1.2 BraTS 2021 Data

The BraTS2021 dataset contains a total of 2000 subjects but now only the official training set are publicly available now which contains 1251 subjects. ¹

Each patient has four modalities in the BraTS dataset, including T1, T1ce, T2, and FLAIR data and the corresponding ground truth labels. Some preprocessing steps have already been taken by BraTS 2021 organizers, such as image registration and skull stripping.

The original dataset contains four types of labels: background (0), necrotic part (1), peritumoral edematous/invaded tissue (2), and enhancing tumor (4). As suggested by the BraTS organizers, labels 1, 2, and 4 should be combined as Whole Tumor (WT) class, labels 1 and 4 should be combined as Tumor Core (TC), and label 2 alone as Enhancing Tumor (ET).

Typically, data from all four modalities are used for normal segmentation tasks with the BraTS dataset. However, in UDA tasks, the model is usually trained on only one modality and does domain adaptation step on another [14, 15]. Therefore, in this project, T1ce is chosen as the source domain and T2 as the target domain. This choice is because not everyone could easily get a T1ce scan as it requires an intrabody injection of contrast medium as described in Section 1.1.1.

¹The official validation set and test set is only open for participants during 30 Jul 2021 - 30 Oct 2021

Unlike the CrossModa 2022 dataset, all subjects in the BraTS2021 dataset have segmentation labels, which makes it possible to test model performance locally. Specifically, this project imitates the dataset structure of CrossModa 2022, and divides the BraTS 2021 training set in an unpaired way as described as follows:

- **Source Training Set:** 500 patients are randomly selected as the source domain training set, only their T1ce data and segmentation labels are used in this set.
- **Target Training Set:** Another 500 patients are randomly selected as the target domain training set, and only their T2 images are used to perform unsupervised training.
- **Validation and Test Set:** Among the remaining 251 samples, 125 are randomly selected as the validation set, and the remaining 126 are used as the test set. The validation and test sets are only used to validate model performance, and their ground truth labels are never provided to the model to learn.

3.2 Evaluation Metrics

3.2.1 Evaluation Metrics for Classification

The commonly used evaluation metrics for classification tasks are calculated from the confusion matrix. For a binary classification problem, its confusion matrix includes the following four categories:

- **TP:** where a sample is predicted as positive class and the prediction is correct.
- **FP:** where a sample is predicted as positive class and the prediction is incorrect.
- **TN:** where a sample is predicted as negative class and the prediction is correct.
- **FN:** where a sample is predicted as negative class and the prediction is incorrect.

Based on these four values, precision, recall and f1 score can be defined as follows:

$$precision = \frac{TP}{TP + FP} \quad (3.1)$$

$$recall = \frac{TP}{TP + FN} \quad (3.2)$$

$$f1 \text{ score} = 2 \times \frac{precision \times recall}{precision + recall} \quad (3.3)$$

It should be noted that the above equation is defined for a single class. For multi-class problem, we can further define micro and macro metrics, such as:

$$micro \text{ precision} = \frac{\sum_{i=1}^n TP_i}{\sum_{i=1}^n TP_i + \sum_{i=1}^n FP_i} \quad (3.4)$$

$$micro\ recall = \frac{\sum_{i=1}^n TP_i}{\sum_{i=1}^n TP_i + \sum_{i=1}^n FN_i} \quad (3.5)$$

$$macro\ precision = \frac{\sum_{i=1}^n precision_i}{n} \quad (3.6)$$

$$macro\ recall = \frac{\sum_{i=1}^n recall_i}{n} \quad (3.7)$$

where the subscript i represents the i th class, and n is the total number of classes.

As there is a severe class imbalance problem in the CrossModa 2022 source domain training set, this project adopts the macro metrics when evaluating the source domain performance because a micro metrics is unreliable as the majority class could easily mislead it.

For the CrossModa 2022 target domain, the ground truth labels are hidden for participants, so it's impossible to calculate precision, recall, and f1 locally. Instead, the participants must upload their predictions to the official website, and then the organiser will provide the Macro Averaged Mean Absolute Error (MA-MAE) [44] metric as feedback. The mathematical definition of MA-MAE is as follows:

$$MA - MAE = \frac{1}{n} \sum_{j=1}^n \frac{1}{n_j} \sum_{x_i \in T_j} |pred_{x_i} - true_{x_i}| \quad (3.8)$$

where n is the total number of classes, n_j is the number of samples with ground truth class j , $x_i \in T_j$ is the i th sample in the test set with a ground truth label j .

3.2.2 Evaluation Metrics for Segmentation

Dice score is a commonly used metric in medical image segmentation tasks. It is a measure of similarity. Its range is between $[0, 1]$. The larger the value, the better the segmentation result. The mathematical definition of dice is as follows:

$$dice = \frac{2 \times (pred \cap true)}{pred \cup true} \quad (3.9)$$

where $pred$ represents the set of predicted values, and $true$ is the set of ground truth values. At the same time, if we regard the segmentation task as a pixel-wise classification task, we can use the confusion matrix to rewrite Equation 3.9 as:

$$dice = \frac{2 \times TP}{FP + 2 \times TP + FN} \quad (3.10)$$

Chapter 4

Methodologies

This chapter will detail the methods used in this project for the UDA classification problem and the UDA segmentation problem. Specifically, for the classification problem, the Cross Domain Transformer (CDTrans) [45] model proposed by Xu, T et al. is adopted, and an innovative Momentum Prototype UDA (MP-UDA) model is proposed for the segmentation problem.

4.1 Classification

CDTrans is a network structure for the UDA classification problem published by Xu et al. [45] this March. It is one of the first attempts to use the transformer idea to solve the UDA task in recent years. It archives state of the art performance on datasets such as office-home, office31, and VisDA2017 [46]. One of the motivations of CDTrans is that the authors found that the transformer structure has a strong tolerance for noisy pseudo labels. This section will first introduce its network architecture and then, its training procedure.

4.1.1 Network Architecture

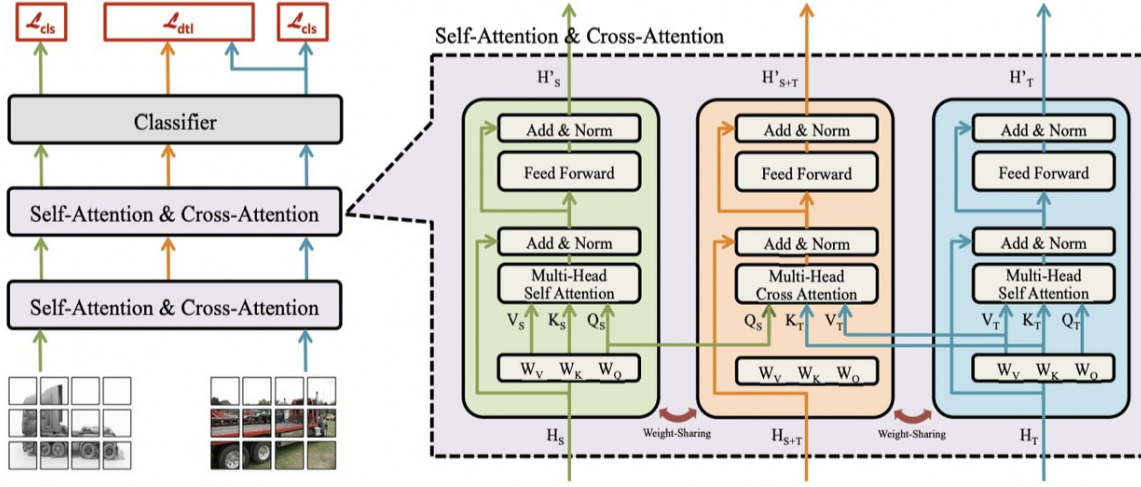


Figure 4.1: Network Architecture of CDTrans

The overall network architecture is shown in Figure 4.1. There are three branches in the network structure and they share the same classifier. During training, the input data is provided in pairs (we will discuss how each pair is constructed in the following sub section). Specifically, the green and blue branches in the figure represent the source branch and the target branch, respectively. They accept images of their respective domains as input to generate *Value*, *Key*, and *Query*, and use self-attention to learn image features. The orange branch in the middle is the source-target branch, which uses the *Query* from the source branch and the *Key* and *Value* from the target branch, and then aligns the feature distributions of the two domains through cross attention.

In theory, the role of the source branch is to use the source ground truth labels in a supervised manner to ensure the model's performance on the source domain and provide appropriate *Query* information for cross attention. The target branch uses pseudo-labels for supervised learning and provides *Key* and *Value* for cross attention. It should be noted that source-target branch does not directly use pseudo-labels for training, but uses distillation loss to guide the target branch training. The relationship between source-target and target branch is similar to teacher and student. Specifically, the equation of distillation loss is as follows:

$$L_{dtl} = \sum_k q_k \log(p_k) \quad (4.1)$$

where q_k and p_k are the output probabilities of a sample being class k from source-target branch and target branch, respectively.

Only the target branch is used for target domain inference.

4.1.2 Training Procedure

The training process of CDTrans consists of 4 steps:

1. Source domain training. Perform supervised training using both source domain data and labels.
2. Two-way labelling. This step focuses on constructing paired data. For each sample s in the source domain \mathcal{S} , find the closest sample t in target domain \mathcal{T} to it through a distance measure $d(\cdot, \cdot)$, and thus obtain P_s . Do the same for the target domain to get P_t :

$$P_s = \{(s, t) | t = \min_k d(f_s, f_k), \forall k \in T, \forall s \in S\} \quad (4.2)$$

$$P_t = \{(s, t) | t = \min_k d(f_t, f_k), \forall t \in T, \forall k \in S\} \quad (4.3)$$

where f represents the extracted feature.

The final paired dataset is $P_s \cup P_t$. For each pair, the pseudo label of the target domain sample is defined as the same as its paired source domain sample.

3. Center Aware Filtering. The pairs obtained in the second step are expected to contain many noisy pseudo-labels, and this step aims to improve the quality of pseudo-labels. Using the source domain model trained in the first step, directly feed target domain samples t to it and obtain the probability distribution δ_t^k , where k means the k th class. After that, k means method is used to find the probability distribution centre of each category:

$$c_k = \frac{\sum_{t \in T} \delta_t^k f_t}{\sum_{t \in T} \delta_t^k} \quad (4.4)$$

After the centre is obtained, the pseudo-label of the target domain sample can be defined by the nearest neighbour classifier:

$$y_t = \arg \min_k d(c_k, f_t) \quad (4.5)$$

After that, we can choose to continue updating the sample centre:

$$c_k' = \frac{\sum_{t \in T} \mathbb{1}(y_t = k) f_t}{\sum_{t \in T} \mathbb{1}(y_t = k)} \quad (4.6)$$

Equation 4.5 and 4.6 can be updated for many iterations, but in the original CDTrans paper, the author only executed it once.

Finally, if the pseudo-label of the target domain sample in a pair is different from that of the source sample, the pair will be discarded. Otherwise it will be kept.

We can see step 2 and 3 follow ‘instance based’ transfer learning approach.

4. UDA. The paired data is fed into the CDTrans network for training. It is worth noting that steps 2 and 3 are repeated after every 5 epochs, with the “source model” being replaced by the current target model.

4.2 Segmentation

For the UDA segmentation problem, this project proposes a new training method called Momentum Prototype UDA (MP-UDA), which is inspired by [38]. The new MP-UDA has the following two advantages:

1. **Source Free.** The source domain data is only used for source domain training. After obtaining the source domain model, there is no need to re-access the source domain data. This characteristic provides convenience for cross-institution cooperation because the two institutions will no longer need to exchange training data, which significantly protects patients' privacy.
2. **Support various network architecture.** The novelty of this method lies in how to use the prototype feature to construct pseudo labels. There are no restrictions or requirements on the network architecture.

An overview of how the data flows in the proposed MP-UDA method is illustrated in Figure 4.2. In this diagram and the following explanation, we assume that there are only two classes in the segmentation task: background and object. However, before we delve into the MP-UDA details, we need to define the following things:

- **Feature:** A feature is the input of the last convolutional layer of a neural network model. Feature is defined in a pixel-wise way. For example, feature is a tensor with shape $(batch\ size, 64, height, width)$ in the actual implementation.
- **Probability:** A probability is the output of the last convolutional layer of a neural network model with softmax activation function applied. Probability is also defined in a pixel-wise way, it has shape $(batch\ size, 2, height, width)$ in the actual implementation.
- **Object Ratio:**

$$Object\ Ratio = \frac{number\ of\ object\ pixels}{number\ of\ all\ pixels} \quad (4.7)$$

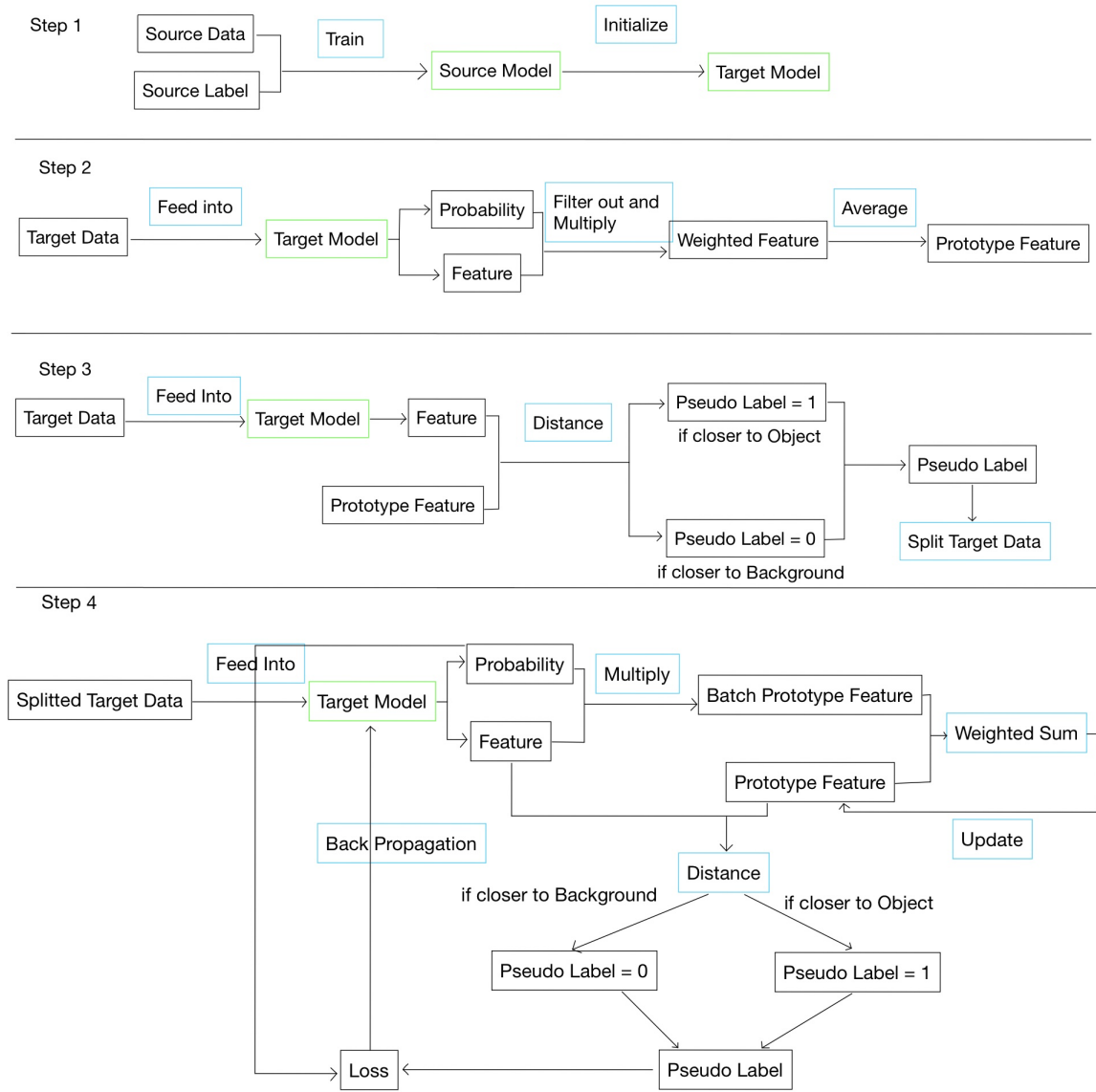


Figure 4.2: Data Flow in MP-UDA. In this figure, black boxes are data, blue boxes are operations and green boxes are model parameters

Specifically, MP-UDA's training process also includes 4 steps:

1. Source model training. Use source data and source labels to train a source model and use this source model to initialise the target model. This is a common step for source-free UDA approaches.
2. Build prototype features. Directly feed target data into the untrained target model, and get two outputs: *feature* and *probability*. We first filter out some pixels if their *probability* is less than *pseudo label threshold*, for the remaining pixels, their *feature* and *probability* are multiplied to get the *weighted feature*. This process is repeated for all target domain training data and finally *prototype feature* is calculated as the average of all *weighted feature*. Note that each class has its own unique *prototype feature*. No training is performed in this

step.

3. Split target domain dataset according to *object ratio*. Directly feed target data into the untrained target model to get *feature*, and use *feature* together with the *prototype feature* to calculate a pixel-wise *distance*. For each pixel, if its *distance* is closer to object, then assign *pseudo label* = 1, otherwise assign *pseudo label* = 0, and use *pseudo label* to split target training set into two parts: one with *object ratio* < *threshold*, the other with *object ratio* ≥ *threshold*. This step is a training trick learned from Coursera to tackle class imbalance problem. No training was performed in this step.
4. Adaptation to the target domain. Feed target data into the target model, and use the same operation as step 2 to calculate a *batch prototype feature*. The *prototype feature* is updated in a momentum manner:

$$\text{prototype feature} = 0.99 \times \text{prototype feature} + 0.01 \times \text{batch prototype feature} \quad (4.8)$$

The updated *prototype feature*, together with the model output *feature* is used to calculate the *pseudo label* in the same way as step 3. Loss is calculated with *pseudo label* and model output *probability*. In this step, loss is a combination of Dice Loss and Entropy Loss as suggested by Liang et al. [13]

Chapter 5

Experiment Results

This chapter will first describe the hardware and software environment in which the code of this project runs, then describes the result achieved in the two tasks and give a detailed analysis of their training process.

5.1 Hardware and Software Environment Setup

In terms of hardware, all experiments in this project are run on an Nvidia RTX 3090 GPU, occupying 24G graphics memory in total.

In terms of software, all the code of this project is based on the PyTorch framework. Among the two tasks, this project uses the official implementation of CDTrans [47], and uses a GitHub BraTS repository [48] as the starting point for the MP-UDA approach. MP-UDA inherits source domain training, data argumentation and some utils functions from the GitHub repository mentioned above.

The python version used is 3.7.10, and the CUDA version is 11.5.

5.2 Classification Result

According to Section 4.1.2, CDTrans training procedure is divided into four steps, but only two involve actual training. These two steps are source domain pre-training and target domain UDA. This section will analyze their training performance and the result in detail.

Choices of hyper-parameters are listed in Table 5.1 for both source domain pre-train and the UDA step.

| CDTrans Hyper-parameters | |
|--------------------------------|-----------------------|
| Network Backbone | DeiT Base |
| Optimiser | SGD |
| Optimiser Weight Decay | 0.0001 |
| Base Learning Rate (Pre Train) | 0.005 |
| Base Learning Rate (UDA) | 0.008 |
| Learning Rate Warm Up Factor | 0.01 |
| Learning Rate Warm Up Epoch | 10 |
| Learning Rate Warm Up Method | Linear |
| Batch Size (Pre Train) | 64 |
| Batch Size (UDA) | 32 |
| Input Size | 256 * 256 * 50 |
| Crop Size | 224 * 224 * 50 |
| Stride Size | 16 * 16 * 50 |
| Random Horizontal Flip | 0.5 |
| Loss (Pre Train) | Softmax Cross Entropy |
| Loss (UDA) | Triplet Loss |

Table 5.1: CDTrans Hyper-parameters

Before source domain pre-train step, we resize all CrossModa 2022 data to three-dimensional tensors of shape (256, 256, 50) to tackle the difficulty that the dataset contains samples of different spatial dimensions. We randomly selected 20% subjects from CrossModa 2022 source domain training set to form source domain validation set.

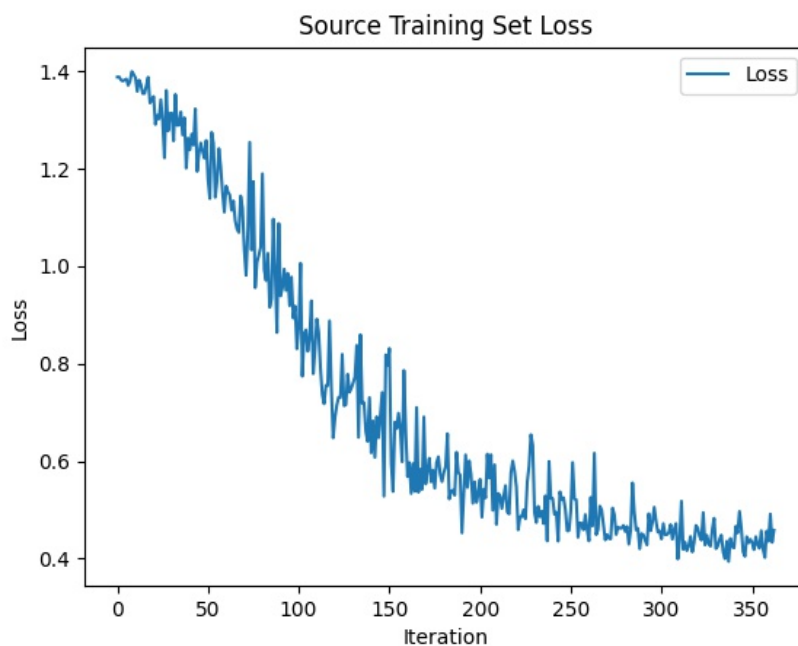


Figure 5.1: CDTrans Pre-train Loss on CrossModa 2022 Source Domain Training Set

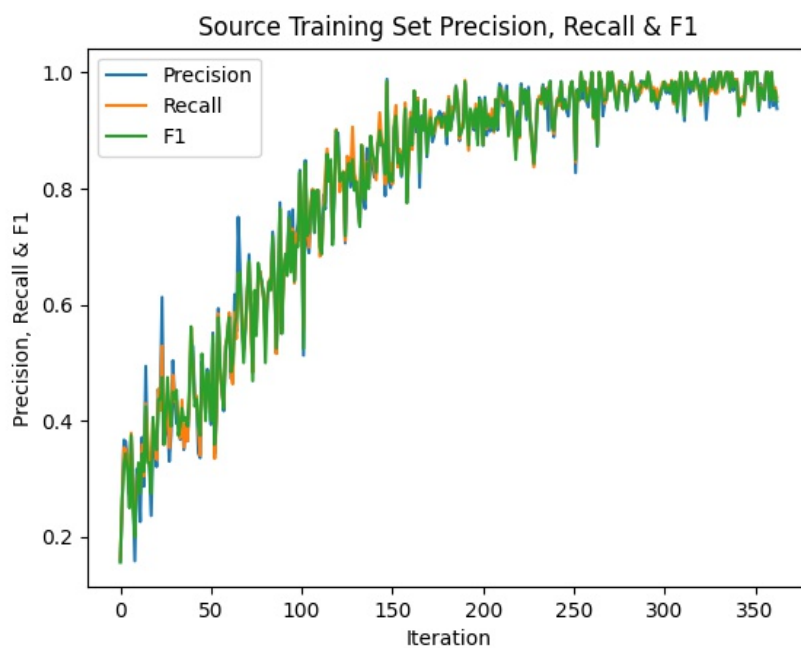


Figure 5.2: Three Classification Metrics of Koos Classification of CDTrans on CrossModa 2022 Source Domain Training Set

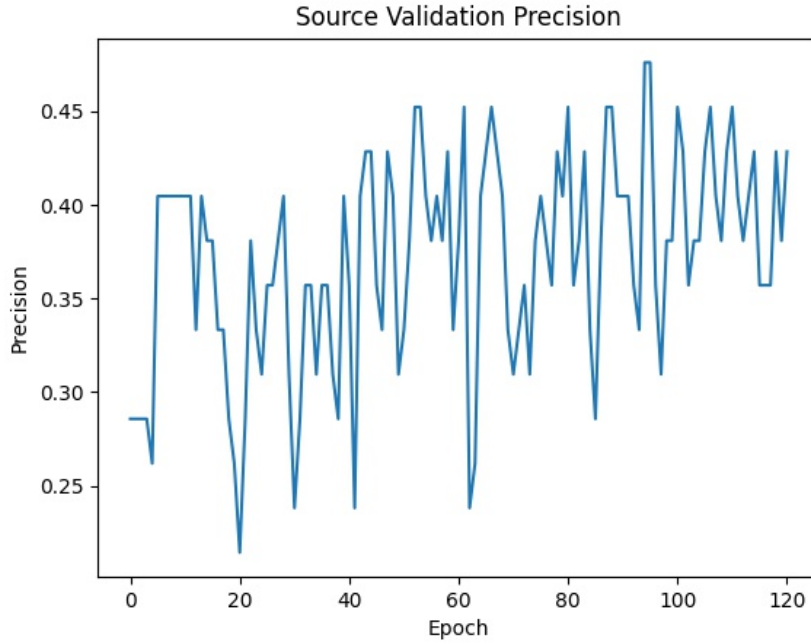


Figure 5.3: Precision of Koos classification of CDTrans on CrossModa 2022 Source Domain Validation Set

Source domain pre-train loss is illustrated in Figure 5.1. It is clear from this figure that the softmax cross entropy loss declines steadily during the training process. Correspondingly, the model performs pretty well on the source domain training set. Figure 5.2 shows how three evaluation metrics evolve during training. It can be seen that macro average recall, macro average precision, and f1 score gradually increase to above 0.9. However, as shown in Figure 5.3, the result is not ideal when testing on the source domain validation set. The highest macro recall can only reach about 0.45. Considering this task only has 4 classes, even a random classifier is expected to achieve an accuracy of 0.25, so it can be concluded that the source domain model simply memorizes the training data it has seen, it learns little task-specific knowledge.

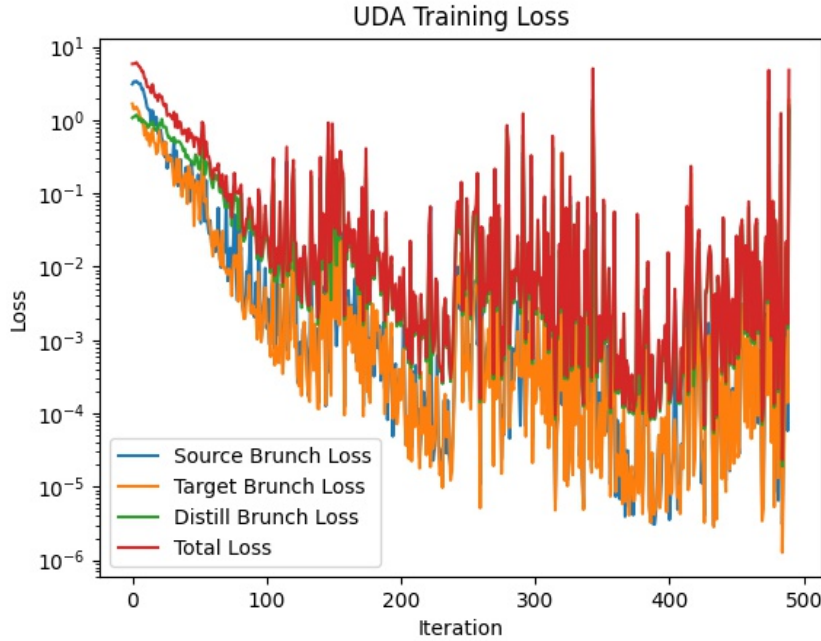


Figure 5.4: CDTrans Loss On CrossModa 2022 Target Domain Training Set

Figure 5.4 shows the loss diagram of the UDA training step of CDTrans. All losses of the three branches are declining steadily in the beginning, but they increase sharply sometimes later. This phenomenon is a result of the fact that both P_s and P_t sets are reconstructed several times during training, resulting in a change of pseudo label distributions. It takes a while for the model to re-adapt to the new training pair set $P = P_s \cup P_t$.























Prediction results are submitted to the official CrossModa website [6] to obtain model performance in the target domain. The only metric they provide is MA-MAE. The official leader-board is shown in Figure 5.5. This project has four submissions, and the best MA-MAE is 0.9888, ranking 12th out of 22 submissions. But unfortunately, the best classification result is obtained by directly applying the source domain model to the target domain. The intention of doing this is to use its result as a baseline. The best result obtained by the UDA step has achieved 1.0411 MA-MAE, ranking 14th on the leader-board.

Based on the above results, it can be concluded that the CDTrans network is unsuitable for this task. The reason for this phenomenon may be because the volume of the VS is too small compared with the whole brain, it is hard for the model to find task-specific information when provided with the whole brain as input, so even the source domain supervised learning cannot achieve ideal performance. Furthermore, in CDTrans steps 2 and 3, the pseudo-labels generated by the source domain model contain a lot of noise, which further impairs target domain learning.

Task 2: Koos classification - Validation phase Leaderboard

Search:

Additional metrics ▾ Show all metrics

| # | ↑↓ | User (Team) | ↑↓ | Created | ↑↓ | Macro-Average Mean Square Error | ↑↓ | Comment | ↑↓ |
|------|----|---|----|--------------|----|---------------------------------|----|--------------|----|
| 1st | |  kathrynwilkins (SJTU_EIEE_2-426Lab) | | 11 Aug. 2022 | | 0.2184 ± | | 4 | |
| 2nd | |  kathrynwilkins (SJTU_EIEE_2-426Lab) | | 9 Aug. 2022 | | 0.2348 ± | | 2 | |
| 2nd | |  kathrynwilkins (SJTU_EIEE_2-426Lab) | | 30 July 2022 | | 0.2348 ± | | 1 | |
| 4th | |  kathrynwilkins (SJTU_EIEE_2-426Lab) | | 10 Aug. 2022 | | 0.3017 ± | | 3 | |
| 5th | |  hanluyi4869 (Super Polymerization) | | 9 Aug. 2022 | | 0.3940 ± | | docker | |
| 6th | |  hanluyi4869 (Super Polymerization) | | 20 July 2022 | | 0.4615 ± | | msf2dfa | |
| 7th | |  hanluyi4869 (Super Polymerization) | | 19 July 2022 | | 0.4874 ± | | msf25dsegf0 | |
| 8th | |  hanluyi4869 (Super Polymerization) | | 16 July 2022 | | 0.6805 ± | | msf25dsegcon | |
| 9th | |  yunzhi.huang.scu@gmail.com (Super Polymerization) | | 16 July 2022 | | 0.8371 ± | | | |
| 10th | |  SKJP | | 9 Aug. 2022 | | 0.8405 ± | | | |
| 11th | |  SKJP | | 3 Aug. 2022 | | 0.9566 ± | | | |
| 12th | |  wenqing.zong98 (CrossModaTeam) | | 20 July 2022 | | 0.9888 ± | | | |
| 13th | |  hwangjeongyong4 (DMCB) | | 9 Aug. 2022 | | 1.0011 ± | | 1 | |
| 14th | |  wenqing.zong98 (CrossModaTeam) | | 24 July 2022 | | 1.0411 ± | | | |
| 14th | |  wenqing.zong98 (CrossModaTeam) | | 23 July 2022 | | 1.0411 ± | | | |
| 16th | |  wenqing.zong98 (CrossModaTeam) | | 25 July 2022 | | 1.0816 ± | | | |
| 17th | |  shmoon (DMCB) | | 2 Aug. 2022 | | 1.0920 ± | | decision | |
| 18th | |  shmoon (DMCB) | | 5 Aug. 2022 | | 1.1083 ± | | | |
| 19th | |  SKJP | | 24 July 2022 | | 1.1328 ± | | | |
| 20th | |  SKJP | | 2 Aug. 2022 | | 1.1497 ± | | | |
| 21st | |  SKJP | | 18 July 2022 | | 1.1580 ± | | | |
| 22nd | |  shmoon (DMCB) | | 3 Aug. 2022 | | 1.2474 ± | | | |

Showing 1 to 22 of 22 entries

Previous 1 Next

Figure 5.5: CrossModa 2022 Official Leader-Board for Classification Task [49]

5.3 Segmentation Result

Unlike the classification task's failure, MP-UDA achieves remarkable achievements for the segmentation task on BraTS2021 dataset. The hyper-parameters used in MP-UDA training are shown in the following Table 5.2:

| MP-UDA Hyper-parameters | |
|--|---|
| Network Backbone | PSPNet with Resnet50 as Encoder |
| Batch Size (Source Domain Pretrain) | 48 |
| Batch Size (Target Domain UDA) | 40 |
| Optimizer | SGD |
| Loss Function (Source Domain Pretrain) | Dice Loss |
| Loss Function (Target Domain UDA) | Dice Loss + Entropy Loss |
| Epoch (Source Domain Pretrain) | 5 |
| Epoch (Target Domain UDA) | 1 |
| Learning Rate (Source Domain Pretrain) | 0.001 |
| Learning Rate (Target Domain UDA) | 0.0005 |
| Dropout Rate | 0.1 |
| Prototype Feature Update Momentum | 0.995 |
| Pseudo Label Probability Thresh | 0.75 |
| Whole Tumor Threshold | 0.05 |
| Whole Tumor Rate | 0.66 |
| Core Tumor Threshold | 0.03 |
| Core Tumor Rate | 0.66 |
| Enhancing Tumor Threshold | 0.02 |
| Enhancing Tumor Rate | 0.7 |
| Random Horizontal Flip | 0.5 |
| Random Vertical Flip | 0.5 |
| Random Affine | Within degrees (-20, 20), translate (0.1, 0.1), scale (0.9, 1.1), and shear (-0.2, 0.2) |
| Elastic Transform | Alpha 720 and Sigma 24 |

Table 5.2: MP-UDA Hyper-parameters and Data Augmentation

Batch size is selected as the maximum value that GPU memory can accept, so it is not a typical ‘power of 2’ choice. Only two label classes are used during each training: the background and the foreground. We will use the term ‘foreground’ and ‘object’ interchangeably. The training dataset is divided into two parts according to *label threshold*: part 1 contains all images whose *object ratio* is greater than *label threshold*, and part 2 contains everything else. When a batch is formed, each image has *label rate* probability of being randomly selected from part 1, and $1 - \text{label rate}$ probability to be randomly selected from part 2. Note that the same batch forming mechanism is used in both source and target domain training, except that the source domain uses ground truth labels to calculate *object ratio*, whereas target domain UDA uses pseudo labels.



Figure 5.6: MP-UDA Dice Coefficient During BraTS 2021 Source Domain Training

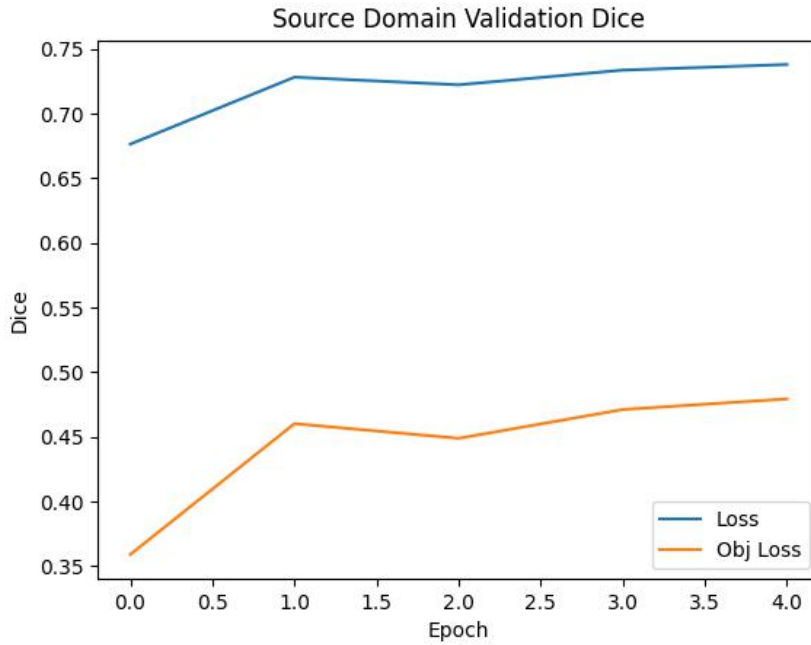


Figure 5.7: MP-UDA Performance on BraTS 2021 Validation Set

Next, we will analyse the performance of source and target domain training in detail. The dataset partition has been described in detail in Section 3.1.2. Similar to CDTrans, the first step of MP-UDA is a regular supervised learning. Dice loss is used

as the loss function, and dice coefficient is used as the evaluation metric to inspect the training process. There is a straightforward quantitative relationship between them: $\text{dice loss} = 1 - \text{dice coefficient}$. Figure 5.6 shows the model's dice coefficient change on the training set for experiment with two classes: background and Whole Tumor. It is clear that both the overall dice and the object dice coefficient show a steady upward trend. Model performance on the validation set is tested after each epoch, and the result is shown in Figure 5.7. The model obtains a dice coefficient of about 0.74 on the validation set and there is no performance drop as the training progresses. As a result, a conclusion can be drawn that the source domain supervised learning is successful. As for the remaining two experiments (background-Core Tumor and background-Enhancing Tumor), the dice coefficient shows a similar behaviour as depicted in the above mentioned figures.

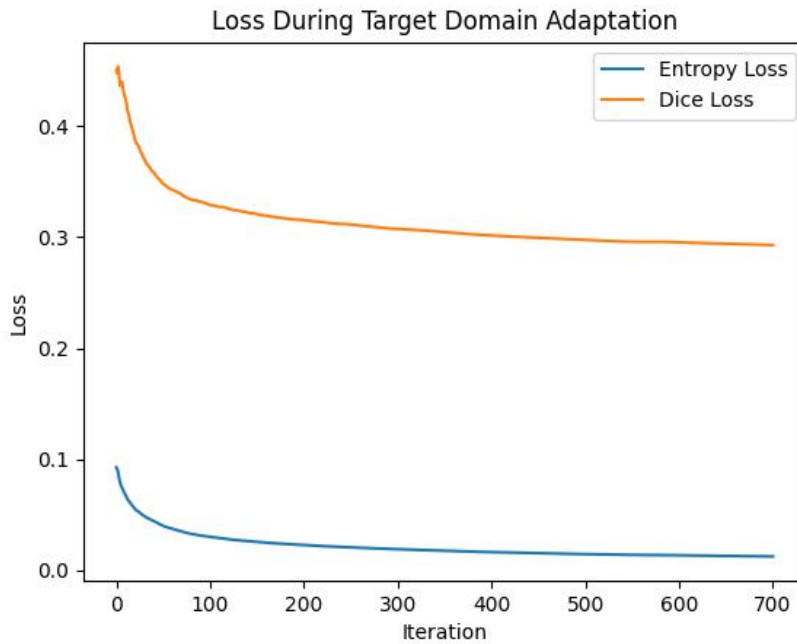


Figure 5.8: MP-UDA Loss On BraTS 2021 Target Domain Training Set

Loss change during target domain UDA training is shown in Figure 5.8. In addition to dice loss, which is calculated by model output and pseudo labels in this case, MP-UDA also adopts [13]’s idea of using an additional entropy loss to help the network converge. As shown in the above figure, both losses steadily decrease during the UDA training process. This indicates that the constructed pseudo labels can indeed benefit the target model to transfer knowledge from source model.

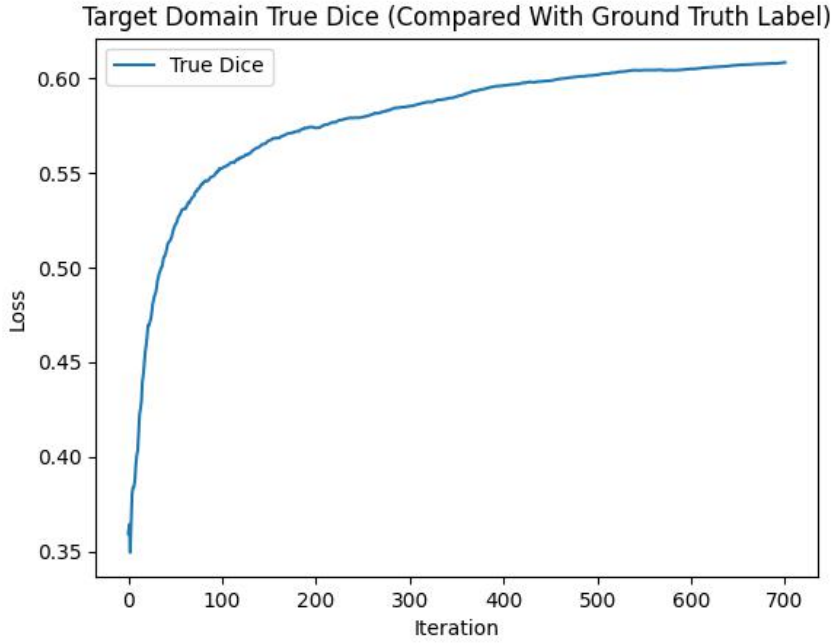


Figure 5.9: MP-UDA True Dice Coefficient On BraTS 2021 Target Domain Training Set

Figure 5.9 depicts the true dice coefficient of the target model which is calculated by model output and ground truth label. It can be seen that the prediction of the target domain model gets better and better as the training proceeds. At the very beginning, the untrained target modal only achieves 0.35 dice coefficient due to severe domain shift, but it quickly completes knowledge transfer and reaches an epoch average of 0.62 dice coefficient after one epoch. It should be pointed out that this true dice is only used as supervision for the target domain training process, and it never participates in back propagation. In other words, the target domain model learns nothing from the ground truth label.

Both Figure 5.8 and 5.9 have been smoothed. Specifically, each point in the graph is the average of all previous data.

In addition to understanding the MP-UDA performance through training process loss change, this project also re-implemented Liu et al.'s BBUDA model [14] as a comparison. BBUDA is the latest UDA model published in this June. The results are shown in Table 5.3. Note the results here are reported on the test set which contains 126 subjects. In the UDA step, we ensure that BBUDA uses the same source domain model and hyper-parameters as MP-UDA where ever possible for fair comparison so any difference is purely caused by different UDA algorithm.

In comparison with BBUDA, our new method achieves similar results as BBUDA in categories Tumor Core and Enhancing Tumor, but it falls behind by a small gap in Whole Tumor category. We additionally compare the performance between MP-UDA and the SOTA [50] method on the BraTS 2021 dataset. It should be pointed out that the task of this SOTA method is not UDA but ordinary supervised learning, and it

uses data from four modalities, so it is expected that the performance of MP-UDA is not as good as that of SOTA.

In addition, we noticed that there is a huge performance gap between re-implemented BBUDA and the original BBUDA paper, so we listed all the modifications made in the re-implementation in Tabel 5.4.

| Model | Whole Tumor Dice | Tumor Core Dice | Enhancing Tumor Dice |
|------------------------|------------------|-----------------|----------------------|
| Source Only | 0.374 | 0.358 | 0.379 |
| MP-UDA | 0.667 | 0.614 | 0.626 |
| BBUDA (re-implemented) | 0.673 | 0.613 | 0.627 |
| BBUDA (original paper) | 0.763 | 0.373 | 0.396 |
| BraTS 2021 SOTA | 0.846 | 0.905 | 0.853 |

Table 5.3: Comparison With BBUDA and BraTS 2021 SOTA

| Modifications Made To The Re-implemented BBUDA | | |
|--|----------------------|---------------------------------|
| | Original BBUDA Paper | Re-implemented BBUDA |
| Network Backbone | 2D Unet | PSPNet with Resnet50 as Encoder |
| Dataset | BraTS2018 | BraTS2021 |
| Source Domain Modality | T1 | T1ce |
| Input Spatial Dimension | 128 * 128 | 240 * 240 |
| Epoch (Source Domain Pretrain) | 100 | 5 |
| Epoch (Target Domain UDA) | 100 | 1 |
| Learning Rate | Not Mentioned | 0.0005 |

Table 5.4: Modifications Made To The Re-implemented BBUDA

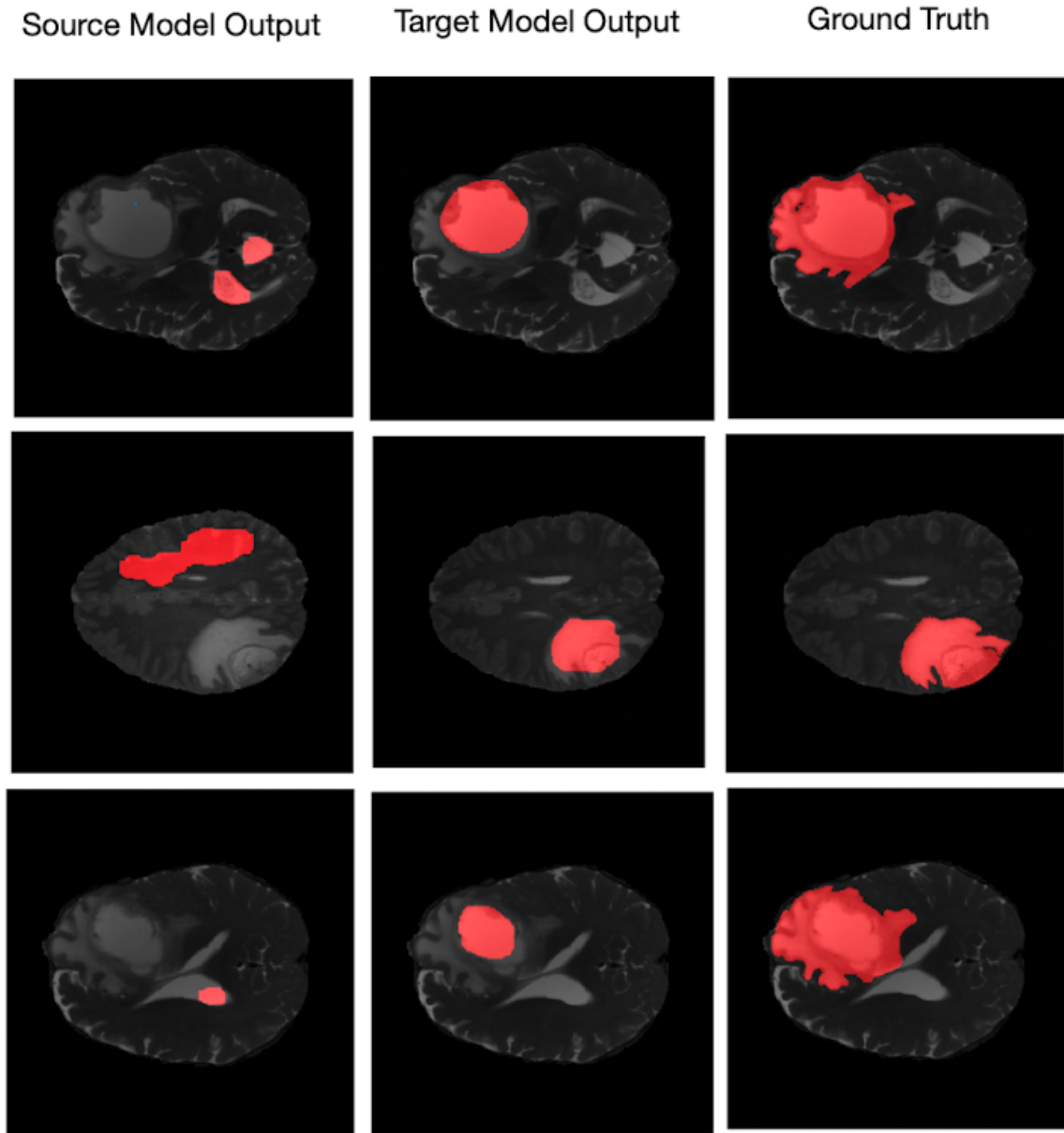


Figure 5.10: Visualisation of MP-UDA Performance

The actual output of MP-UDA is shown in the Figure 5.10. The left column is the prediction result obtained by directly applying the source domain model to the target domain data, the middle column is the model output after MP-UDA training, and the far right is Ground Truth. As can be seen from the figure, the prediction of the source domain model is almost completely staggered from Ground Truth. Although MP-UDA learns from this bad source domain model, it can correctly find the general location of the tumor after UDA training. However, it should still be noted that the output of MP-UDA seems to be conservative and cannot cover the entire ground truth label.

Chapter 6

Conclusions and Future Work

This chapter will give an overall summary of the project and analyse possible future research ideas.

6.1 Conclusions

Deep learning has achieved remarkable achievements in medical image analysis tasks, but in actual practical applications, ground truth labels are not always available, which makes supervised training impossible. Also, domain shift prevents us directly applying a trained model to another dataset/modality/institution. Therefore, this project attempts to adopt UDA approach to solve the shortcomings mentioned above in two common types of medical image analysis tasks: classification and segmentation.

In the classification part, we use the CDTrans [45] network, which is one of the first attempts to solve UDA via using the transformer idea, but the result is not ideal. This is partially due to the lack of the corresponding labels of the validation set, which makes analysing model performance locally impossible. Another reason is: the selected CDTrans network may not be suitable for the CrossModa 2022 dataset [6]. Compared to the whole brain, only a tiny portion determines VS grade. CDTrans cannot learn in this extreme situation, so even source domain supervised learning cannot achieve an ideal result.

In the segmentation part, inspired by [38] et al.'s approach of using prototype features, we propose a novel Momentum Prototype UDA (MP-UDA) based on momentum update. This method uses prototype features to construct pseudo-labels to help target domain models to improve performance. Based on the result, we can say MP-UDA fulfils this project goal. Also, compared with the latest model proposed by Liu et al. [14] this June, MP-UDA achieves a similar performance.

6.2 Future Work

Based on the success and failure of this project, we propose the following future research ideas:

- Using the Transformer network architecture to solve the UDA problem is a new idea. As a pioneer, CDTrans has achieved state-of-the-art results in natural image classification tasks, but its performance on small targets, such as VS in the brain, is not ideal. Therefore, one possible UDA research direction in the future is to classify small objects via transformer architecture.
- MP-UDA has achieved good results on the BraTS2021 segmentation task. Still, it should be noted that cross-domain is only one possible scenario in practical applications, and UDA may also be applied in cross-institutional cooperation. In the latter situation, the data privacy of each organization will become a huge problem that has to be considered. Although MP-UDA is a source-free method that can partially solve the problem of data leakage, it is still a victim of Model Inversion attack [51] because it needs to share the source domain model. Therefore, another possible future research idea is to defend against model inversion before sharing the source domain model to completely prevent data leakage.
- When compared with the BraTS2021 SOTA method, MP-UDA has a clear performance gap, although this may be because SOTA uses four modalities of data for supervised learning whereas MP-UDA can only use one modality of data and aims for UDA problem.

Chapter 7

Ethical Considerations

This chapter will introduce the ethical problems encountered in the project from three aspects: data, method and code.

7.1 Data Ethics

The Imperial Ethics Checklist is included in Figure A.1 and Figure A.2 in Appendix A. From the data perspective, the biggest ethical issue in this project is patient privacy. We will describe how this project protects the privacy of patient data in two sections.

7.1.1 CrossModa 2022 Dataset Privacy Protection

Although the data set provided by CrossModa 2022 has not been pre-processed by Skull Stripping, but the London data blurred the patient's face information (as shown in Figure 7.1), while the Tilburg data directly cropped out patients' face (as shown in Figure 7.2). As a result, it is impossible to restore the patient's facial information through the officially provided dataset. In addition, there is no other information in this dataset except the 3D image data and the corresponding labels of the source domain, so the anonymity and privacy of the patient is protected.

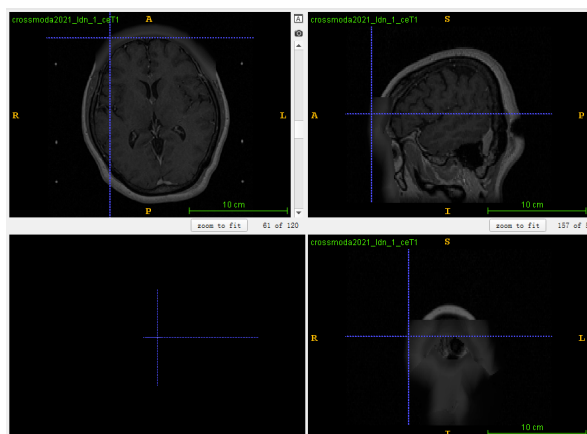


Figure 7.1: CrossModa 2022 London Data Blurs Patients' Facial Information

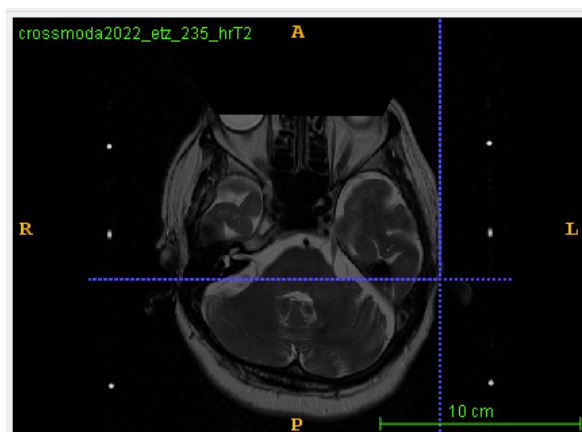


Figure 7.2: CrossModa 2022 Tilburg Data Crops Patients' Facial Information

7.1.2 BraTS 2021 Dataset Privacy Protection

In the BraTS 2021 data, the organizer has performed Skull Stripping on 3D images of all patients (as shown in Figure 7.3). The dataset only contains brain images and ground truth segmentation labels, and there is no face information, so the anonymity and privacy of patients are protected.

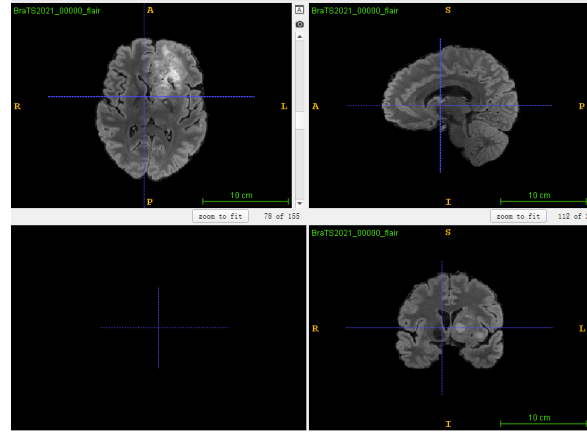


Figure 7.3: BraTS 2021 Data is Skull Stripped

7.2 Method Ethics

The MP-UDA is designed to be source-free, but not black box. Although Black Box can better prevent Model Inversion (MI) attacks by encapsulating the source domain model into an API, this method inevitably feeds the target domain data into the source domain model, and attackers can easily launch a Man In The Middle (MITM) attack to steal target domain data.

And, if we consider the worst scenario, MI attacks are far less harmful than MITM, because MI cannot reconstruct 100% accurate source domain training data. However, once MITM succeeds, it will get utterly accurate target domain data. That forms the motivation for avoiding Black Box design.

7.3 Code Ethics

The code of this project is developed based on two open source repositories [47, 48]. In the original repositories, they all use the MIT license, so modifications to the code and use it for private purposes are allowed, provided that the modified code must also use the MIT licence.

In addition, although MP-UDA solves the problems of domain shift and source-free, it cannot be directly applied to practical applications of T1ce to T2 transfer learning due to the obvious performance gap between MP-UDA and BraTS2021 SOTA. However, we believe that if we use more modality to train the source domain model, it can help the target domain model to achieve better performance.

Bibliography

- [1] Drzezo. *Introduction to Medical Imaging — Radiology Key*; [Accessed 30th Aug 2022]. Available from: <https://radiologykey.com/introduction-to-medical-imaging/>. pages 1
- [2] UK CR. *MRI scan — Tests and scans*; [Accessed 30th Aug 2022]. Available from: <https://www.cancerresearchuk.org/about-cancer/cancer-in-general/tests/mri-scan>. pages 1
- [3] My-MS. *MRI Basics*; [Accessed 30th Aug 2022]. Available from: https://my-ms.org/mri_basics.htm. pages 2
- [4] Koos WT, Day JD, Matula C, Levy DI. *Neurotopographic considerations in the microsurgical treatment of small acoustic neurinomas*. American Association of Neurological Surgeons. 1998;88:506-12. pages 3, 14
- [5] NHS. *Acoustic neuroma (vestibular schwannoma)*; [Accessed 30th Aug 2022]. Available from: <https://www.nhs.uk/conditions/acoustic-neuroma/>. pages 3
- [6] crossMoDA; [Accessed 30th Aug 2022]. Available from: <https://crossmoda-challenge.ml/>. pages 3, 13, 29, 37
- [7] Roan S. Stereotactic Therapy Best for Brain Metastases and More News From ASTRO — Everyday Health; [Accessed on 30 Aug 2022]. Available from: <https://www.everydayhealth.com/cancer/stereotactic-therapy-is-best-for-brain-metastases-and-more-cancer-news-from-day>. pages 4
- [8] CLINIC M. *Brain metastases - Diagnosis and treatment - Mayo Clinic*; [Accessed 30th Aug 2022]. Available from: <https://www.mayoclinic.org/diseases-conditions/brain-metastases/diagnosis-treatment/drc-20350140>. pages 4
- [9] Shinde A, Akhavan D, Sedrak M, Glaser S, Amini A. *Shifting paradigms: whole brain radiation therapy versus stereotactic radiosurgery for brain metastases*. CNS Oncology. 2019 3;8:CNS27. Available from: [/pmc/articles/PMC6499015/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6499015/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6499015/). pages 4

-
- [10] Wang M, Deng W. *Deep Visual Domain Adaptation: A Survey*. Neurocomputing. 2018 2;312:135-53. Available from: <http://arxiv.org/abs/1802.03601><http://dx.doi.org/10.1016/j.neucom.2018.05.083>. pages 7
- [11] Csurka G, Volpi R, Chidlovskii B. *Unsupervised Domain Adaptation for Semantic Image Segmentation: a Comprehensive Survey*. 2021 12. Available from: <https://arxiv.org/abs/2112.03241v1>. pages 7
- [12] Zhao S, Li B, Reed C, Xu P, Keutzer K. *Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey*. 2020 2. Available from: <https://arxiv.org/abs/2002.12169v1>. pages 7
- [13] Liang J, Hu D, Feng J. *Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation*. 37th International Conference on Machine Learning, ICML 2020. 2020 2;PartF168147-8:5984-95. Available from: <https://arxiv.org/abs/2002.08546v6>. pages 8, 11, 12, 24, 33
- [14] Liu X, Yoo C, Xing F, Kuo CCJ, Fakhri GE, Kang JW, et al. *Unsupervised Black-Box Model Domain Adaptation for Brain Tumor Segmentation*. Frontiers in Neuroscience. 2022 6;0:341. pages 8, 12, 16, 34, 37
- [15] Zhang H, Zhang Y, Jia K, Zhang L. *Unsupervised Domain Adaptation of Black-Box Source Models*. 2021 1. Available from: <https://arxiv.org/abs/2101.02839v2>. pages 8, 12, 16
- [16] Khan MNA, Heisterkamp DR. *Adapting instance weights for unsupervised domain adaptation using quadratic mutual information and subspace learning*. Proceedings - International Conference on Pattern Recognition. 2016 1;0:1560-5. pages 9
- [17] Tan B, Zhang Y, Pan SJ, Yang Q. *Distant Domain Transfer Learning*. 2017. Available from: <http://qzone.qq.com>. pages 9
- [18] Pan SJ, Tsang IW, Kwok JT, Yang Q. *Domain adaptation via transfer component analysis*. IEEE Transactions on Neural Networks. 2011 2;22:199-210. pages 9
- [19] Duan L, Tsang IW, Xu D. *Domain transfer multiple kernel learning*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012;34:465-79. pages 9
- [20] Long M, Wang J, Ding G, Sun J, Yu PS. *Transfer Joint Matching for Unsupervised Domain Adaptation*. 2014. pages 9
- [21] Zhang J, Li W, Ogunbona P. *Joint Geometrical and Statistical Alignment for Visual Domain Adaptation*. 2017. pages 9
- [22] Yosinski J, Clune J, Bengio Y, Lipson H. *How transferable are features in deep neural networks?* Advances in Neural Information Processing Systems. 2014 11;4:3320-8. Available from: <https://arxiv.org/abs/1411.1792v1>. pages 10
-

-
- [23] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. *Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?* IEEE Transactions on Medical Imaging. 2017 6;35:1299-312. Available from: <http://arxiv.org/abs/1706.00712><http://dx.doi.org/10.1109/TMI.2016.2535302>. pages 10
- [24] Ghifary M, Kleijn WB, Zhang M. *Domain Adaptive Neural Networks for Object Recognition*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014 9;8862:898-904. Available from: <https://arxiv.org/abs/1409.6041v1>. pages 10
- [25] Tzeng E, Hoffman J, Zhang N, Saenko K, Darrell T. *Deep Domain Confusion: Maximizing for Domain Invariance*. 2014 12. Available from: <https://arxiv.org/abs/1412.3474v1>. pages 10
- [26] Long M, Cao Y, Wang J, Jordan MI. *Learning Transferable Features with Deep Adaptation Networks*. 32nd International Conference on Machine Learning, ICML 2015. 2015 2;1:97-105. Available from: <https://arxiv.org/abs/1502.02791v2>. pages 10
- [27] Long M, Zhu H, Wang J, Jordan MI. *Deep Transfer Learning with Joint Adaptation Networks*. 34th International Conference on Machine Learning, ICML 2017. 2016 5;5:3470-9. Available from: <https://arxiv.org/abs/1605.06636v2>. pages 10
- [28] Pan SJ, Yang Q. *A survey on transfer learning*. IEEE Transactions on Knowledge and Data Engineering. 2010;22:1345-59. pages 10
- [29] Mihalkova L, Huynh T, Mooney RJ. *Mapping and Revising Markov Logic Networks for Transfer Learning*. 2007:608-14. Available from: www.aaai.org. pages 10
- [30] Davis J, Domingos P. *Deep Transfer via Second-Order Markov Logic*. 2009. pages 10
- [31] Weiss K, Khoshgoftaar TM, Wang DD. *A survey of transfer learning*. Journal of Big Data. 2016 12;3:1-40. Available from: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6>. pages 10
- [32] Li F, Pan SJ, Jin O, Yang Q, Zhu X. *Cross-Domain Co-Extraction of Sentiment and Topic Lexicons*. 2012:8-14. pages 10
- [33] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. *A Comprehensive Survey on Transfer Learning*. Proceedings of the IEEE. 2019 11;109:43-76. Available from: <https://arxiv.org/abs/1911.02685v3>. pages 10
- [34] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. *Domain-Adversarial Training of Neural Networks*. Advances in Computer Vision and Pattern Recognition. 2015 5;17:189-209. Available from: <https://arxiv.org/abs/1505.07818v4>. pages 11
-

- [35] Jiang J, Hu YC, Tyagi N, Zhang P, Rimner A, Mageras GS, et al. *Tumor-aware, Adversarial Domain Adaptation from CT to MRI for Lung Cancer Segmentation*. Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 2018;11071:777. Available from: [/pmc/articles/PMC6169798//pmc/articles/PMC6169798/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6169798/](https://pubmed.ncbi.nlm.nih.gov/311071777/). pages 11
- [36] Cai J, Zhang Z, Cui L, Zheng Y, Yang L. *Towards cross-modal organ translation and segmentation: A cycle- and shape-consistent generative adversarial network*. Medical Image Analysis. 2019 2;52:174-84. pages 11
- [37] Sankaranarayanan S, Balaji Y, Jain A, Lim SN, Chellappa R. *Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2017 11:3752-61. Available from: <https://arxiv.org/abs/1711.06969v2>. pages 11
- [38] Chen C, Liu Q, Jin Y, Dou Q, Heng PA. *Source-Free Domain Adaptive Fundus Image Segmentation with Denoised Pseudo-Labeling*. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021 9;12905 LNCS:225-35. Available from: <https://arxiv.org/abs/2109.09735v1>. pages 12, 22, 37
- [39] Liu X, Xing F, Yang C, Fakhri GE, Woo J. *Adapting Off-the-Shelf Source Segmenter for Target Medical Image Segmentation*. Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention. 2021;12902:549. Available from: [/pmc/articles/PMC8562716//pmc/articles/PMC8562716/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8562716/](https://pubmed.ncbi.nlm.nih.gov/3562716/). pages 12
- [40] Yang Y, Soatto S. *FDA: Fourier Domain Adaptation for Semantic Segmentation*. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020 4:4084-94. Available from: <https://arxiv.org/abs/2004.05498v1>. pages 12
- [41] Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. *Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features*. Scientific data. 2017 9;4. Available from: <https://pubmed.ncbi.nlm.nih.gov/28872634/>. pages 13
- [42] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*. IEEE transactions on medical imaging. 2015 10;34:1993-2024. Available from: <https://pubmed.ncbi.nlm.nih.gov/25494501/>. pages 13
- [43] Baid U, Ghodasara S, Mohan S, Bilello M, Calabrese E, Colak E, et al. *The*

- RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*. 2021 7. Available from: <https://arxiv.org/abs/2107.02314v2>. pages 13
- [44] Kujawa A, Dorent R, Connor S, Oviedova A, Okasha M, Grishchuk D, et al. *Automated Koos Classification of Vestibular Schwannoma*. *Frontiers in Radiology*. 2022 3;0:4. pages 14, 15, 18
- [45] Xu T, Chen W, Wang P, Wang F, Li H, Jin R. *CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation*. 2021 9. Available from: <https://arxiv.org/abs/2109.06165v4>. pages 19, 37
- [46] *CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation — Papers With Code*; [Accessed 30th Aug 2022]. Available from: <https://paperswithcode.com/paper/cdtrans-cross-domain-transformer-for>. pages 19
- [47] *CDTrans/CDTrans: [ICLR2022] CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation*; [Accessed 30th Aug 2022]. Available from: <https://github.com/CDTrans/CDTrans>. pages 25, 41
- [48] *cv-lee/BraTs: PyTorch Keras implementation for BraTs (Brain Tumor Segmentation)*; [Accessed 30th Aug 2022]. Available from: <https://github.com/cv-lee/BraTs>. pages 25, 41
- [49] *Leaderboard - Grand Challenge*; [Accessed on 30th Aug 2022]. Available from: <https://crossmoda2022.grand-challenge.org/evaluation/task-2-koos-classification-validation-phase/leaderboard/>. pages 30
- [50] Peiris H, Chen Z, Egan G, Harandi M. *Reciprocal Adversarial Learning for Brain Tumor Segmentation: A Solution to BraTS Challenge 2021 Segmentation Task*. 2022 1. Available from: <https://arxiv.org/abs/2201.03777v1>. pages 34
- [51] Apple QW, Apple DK. *Reconstructing Training Data from Diverse ML Models by Ensemble Inversion*. 2022. pages 38

Appendix A

Imperial Ethics Checklist

| | Yes | No |
|---|-----|-----|
| Section 1: HUMAN EMBRYOS/FOETUSES | | |
| Does your project involve Human Embryonic Stem Cells? | | ✓ |
| Does your project involve the use of human embryos? | | ✓ |
| Does your project involve the use of human foetal tissues / cells? | | ✓ |
| Section 2: HUMANS | | |
| Does your project involve human participants? | ✓ | |
| Section 3: HUMAN CELLS / TISSUES | | |
| Does your project involve human cells or tissues? (Other than from "Human Embryos/Foetuses" i.e. Section 1)? | | ✓ |
| Section 4: PROTECTION OF PERSONAL DATA | | |
| Does your project involve personal data collection and/or processing? | ✓ | |
| Does it involve the collection and/or processing of sensitive personal data (e.g. health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction)? | ✓ | |
| Does it involve processing of genetic information? | | ✓ |
| Does it involve tracking or observation of participants? It should be noted that this issue is not limited to surveillance or localization data. It also applies to Wan data such as IP address, MACs, cookies etc. | | ✓ |
| Does your project involve further processing of previously collected personal data (secondary use)? For example Does your project involve merging existing data sets? | ✓ | |
| Section 5: ANIMALS | | |
| Does your project involve animals? | | ✓ |
| Section 6: DEVELOPING COUNTRIES | | |
| Does your project involve developing countries? | | ✓ |
| If your project involves low and/or lower-middle income countries, are any benefit-sharing actions planned? | N/A | N/A |
| Could the situation in the country put the individuals taking part in the project at risk? | N/A | N/A |
| Section 7: ENVIRONMENTAL PROTECTION AND SAFETY | | |
| Does your project involve the use of elements that may cause harm to the environment, animals or plants? | | ✓ |
| Does your project deal with endangered fauna and/or flora /protected areas? | | ✓ |
| Does your project involve the use of elements that may cause harm to humans, including project staff? | | ✓ |
| Does your project involve other harmful materials or equipment, e.g. high-powered laser systems? | | ✓ |
| Section 8: DUAL USE | | |
| Does your project have the potential for military applications? | | ✓ |
| Does your project have an exclusive civilian application focus? | ✓ | |
| Will your project use or produce goods or information that will require export licenses in accordance with legislation on dual use items? | | ✓ |
| Does your project affect current standards in military ethics – e.g., global ban on weapons of mass destruction, issues of proportionality, discrimination of combatants and accountability in drone and autonomous robotics developments, incendiary or laser weapons? | | ✓ |
| Section 9: MISUSE | | |
| Does your project have the potential for malevolent/criminal/terrorist abuse? | | ✓ |

Figure A.1: Imperial Ethical Checklist Part 1

| | | |
|---|---|---|
| Does your project involve information on/or the use of biological-, chemical-, nuclear/radiological-security sensitive materials and explosives, and means of their delivery? | | ✓ |
| Does your project involve the development of technologies or the creation of information that could have severe negative impacts on human rights standards (e.g. privacy, stigmatization, discrimination), if misapplied? | | ✓ |
| Does your project have the potential for terrorist or criminal abuse e.g. infrastructural vulnerability studies, cybersecurity related project? | | ✓ |
| SECTION 10: LEGAL ISSUES | | |
| Will your project use or produce software for which there are copyright licensing implications? | | ✓ |
| Will your project use or produce goods or information for which there are data protection, or other legal implications? | ✓ | |
| SECTION 11: OTHER ETHICS ISSUES | | |
| Are there any other ethics issues that should be taken into consideration? | ✓ | |

Figure A.2: Imperial Ethical Checklist Part 2