Investigating the Effect of Annotation Styles on the Generalizability of Medical Deep Learning Algorithms

Jillian Cardinell
Department of Electrical and Computer Engineering
McGill University, Montreal
August, 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Master of Science

Abstract

In recent years, supervised deep learning networks have achieved state-of-the-art results in many public medical segmentation challenges. In spite of their success on isolated datasets and challenges, deep learning networks have yet to be widely adapted in the clinic due to practical and generalizability concerns. Poor generalizability in medical deep learning networks is especially detrimental as implementing such a model in the clinic can lead to unreliable predictions, resulting in potentially severe medical consequences. Failure to generalize is often attributed to differences in the population distribution or imaging space of the new data. In many instances, that is not the only cause for concern. Many researchers have overlooked the potential impact of complex variations in the ground-truth annotations. Ground-truth annotations can vary due to a variety of factors including rater biases, differences in semi-automated labelling assistance software, and differences in clinical goals thus resulting in different *annotation styles*. These challenges are particularly abundant in pathological segmentation tasks where "ground-truth" is significantly more subjective. As a result, even the most capable models may show drastic performance drops when applied to a dataset with an incompatible annotation style, making them appear non-transferable or inadequate. In this thesis, we demonstrate the impact of annotation styles on deep learning networks and propose a simple method to manage them. We leverage 8 different large, proprietary Multiple Sclerosis (MS) clinical trial datasets with T2 lesion segmentations. We utilise a simple, in-line style-adapting mechanism, Conditional Instance Normalization (CIN), to model annotation styles across our datasets. We present a series of experiments comparing these models

to several baselines to investigate the impact of disease phenotype on annotation style, and to demonstrate effect of annotation styles on generalizability. We then propose an analysis mechanism based on CIN to identify similar annotation styles, permitting effective dataset aggregation. Lastly, we study approaches to fine-tune an existing network to new annotation styles for sample-efficient continual learning strategies. The results and methods of this thesis can serve as a reference to other researchers on how to discover and manage potential annotation style shifts in their own datasets.

Abrégé

Ces dernières années, les réseaux d'apprentissage profond supervisés ont obtenu des résultats de pointe dans de nombreux défis de segmentation médicale publique. En dépit de leur succès sur des ensembles de données et des défis isolés, les réseaux d'apprentissage profond n'ont pas encore été largement adaptés à la clinique en raison de problèmes pratiques et de généralisabilité. Une mauvaise généralisation des réseaux d'apprentissage profond dans le domaine médical est particulièrement désavantageux, car la mise en œuvre d'un tel modèle en clinique peut engendrer des prédictions peu fiables, entraînant des conséquences médicales potentiellement graves. L'échec de la généralisation est souvent attribué à des différences dans la distribution de la population ou l'espace d'imagerie des nouvelles données. Dans de nombreux cas, ce n'est pas la seule cause d'inquiétude. De nombreux chercheurs ont négligé l'impact potentiel des variations complexes des annotations de base. Les annotations de base peuvent varier en raison de divers facteurs, notamment les biais des évaluateurs, les différences entre les logiciels d'aide à l'étiquetage semi-automatique et les différences entre les objectifs cliniques, ce qui entraîne des styles d'annotation différents. Ces défis sont particulièrement nombreux dans les tâches de segmentation pathologique où la "vérité de base" est beaucoup plus subjective. En conséquence, même les modèles les plus performants peuvent montrer des baisses de performance drastiques lorsqu'ils sont appliqués à un ensemble de données avec un style d'annotation incompatible, ce qui les fait paraître non transférables ou inadéquats. Dans cette thèse, nous démontrons l'impact des styles d'annotation sur les réseaux d'apprentissage profond et proposons une méthode simple pour les gérer. Nous

exploitons 8 grands ensembles différents de données propriétaires d'essais cliniques sur la sclérose en plaques (SEP) avec des segmentations de lésions T2. Nous utilisons un mécanisme simple d'adaptation du style intégré, la normalisation conditionnelle d'instance (CIN), pour modéliser les styles d'annotation dans nos ensembles de données. Nous présentons une série d'expériences comparant ces modèles à plusieurs bases de référence pour étudier l'impact du phénotype de la maladie sur le style d'annotation, et pour démontrer l'effet des styles d'annotation sur la généralisabilité. Nous proposons ensuite un mécanisme d'analyse basé sur le CIN pour identifier les styles d'annotation similaires, permettant une agrégation efficace des ensembles de données. Enfin, nous étudions des approches permettant d'adapter un réseau existant à de nouveaux styles d'annotation pour des stratégies d'apprentissage continu efficaces en termes d'échantillons. Les résultats et les méthodes de cette thèse peuvent servir de référence à d'autres chercheurs sur la façon de découvrir et de gérer les changements potentiels de style d'annotation dans leurs propres ensembles de données.

Contribution of Authors

- NICHYPORUK, B., CARDINELL, J., SZETO, J., MEHTA, R., TSAFTARIS, S., ARNOLD, D. L., AND ARBEL, T. Cohort bias adaptation in aggregated datasets for lesion segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health* (2021), Springer International Publishing, pp. 101–111

B. Nichyporuk lead this publication. I took on the role of assisting with literature research and experimental design. I also contributed to the writing of this paper. Although results and experiments from this paper are not directly presented in this thesis, this paper was instrumental in laying the groundwork for the methods presented in this thesis. This work was then extended in the following publication:

- Under Review: NICHYPORUK*, B., CARDINELL*, J., SZETO, J., MEHTA, R., FALET, J.-P. R., ARNOLD, D. L., TSAFTARIS, S., AND ARBEL, T. Rethinking generalization: The impact of annotation style on medical image segmentation. *Under Review at: The Journal of Machine Learning for Biomedical Imaging* (2022)

B. Nichyporuk and myself contributed equally to this extension publication. My main contributions to this paper were the experiments and results presented in Chapter 5.1 and Chapter 6. I lead the the design, development and execution of these experiments with consult and collaboration with B. Nichyporuk. Certain sections of this paper were omitted from this thesis as they were primarily B. Nichyporuk's work. D.L. Arnold, S. Tsaf-

taris, and T. Arbel were all supervisors of this paper and corresponding research project. All other authors assisted in writing, presentation, or consultation.

All other chapters in this thesis are solely my work, with some guidance by my supervisor, T. Arbel, and the lab Research Scientist, B. Nichyporuk.

Acknowledgements

I would first like to thank my parents, Brett and Karen, for their continued support and love during my entire academic career. Without them, I would not have been able to accomplish anything. I would also like to thank my partner, Liam, for taking care of me during these past few years. I'd also like to thank my friends Michelle, Shireen, Nhu, and Chelsea for their endless emotional support.

Furthermore, I'd like to thank my supervisor, Prof. Tal Arbel, for her valuable insights and guidance throughout my thesis. I'd also like to thank the PVG Research Scientist, Brennan Nichyporuk, for his consistent help in both technical and research matters. Lastly, I'd like to thank Justin Szeto, Kirill Vasilevski, and Eric Zimmermann for their excellent work on the pipeline that this entire thesis depended on.

I am grateful to the International Progressive MS Alliance for supporting this work (grant number: PA-1412-02420), and to the companies who generously provided the clinical trial data that made it possible: Biogen, BioMS, MedDay, Novartis, Roche / Genentech, and Teva. Funding was also provided by the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs.

Table of Contents

List of Figures

List of Tables

List of Acronyms

DL	Deep Learning
MS	Multiple Sclerosis
RRMS	Relapsing Remitting Multiple Sclerosis
PPMS	Primary Progressive Multiple Sclerosis
SPMS	Secondary Progressive Multiple Sclerosis
DAWM	Diffusely Abnormal White Matter
CNS	Central Nervous System
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
FLAIR	Fluid Attenuated Inverse Recovery
CIN	Conditional Instance Normalization
IN	Instance Normalization
BN	Batch Normalization
CNN	Convolutional neural Network
PR-AUC	Precision-Recall Area Under Curve
FL	Federated Learning
DA	Domain Adaptation

Chapter 1

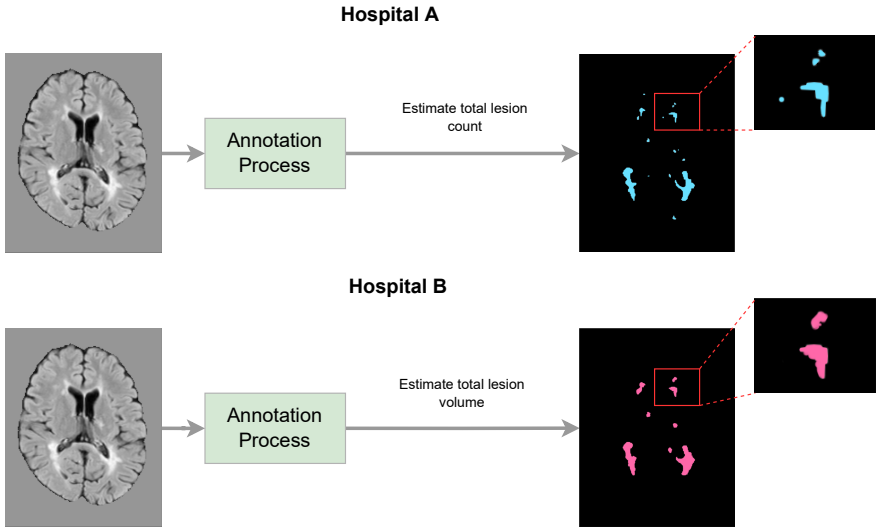
Introduction

Deep learning (DL) methods have greatly advanced the field of computer vision in recent years, and their potential for improving medical image analysis has been explored with great enthusiasm. DL approaches have shown especially beneficial impacts on automated medical segmentation tasks such as cancer segmentation [5], breast calcification segmentation [84], and anatomy segmentation [54]. DL has become state-of-the-art for these segmentation tasks due to their ability to model complex tasks without feature engineering while also obtaining excellent performance. Despite the promising performance of DL approaches on medical imaging tasks, there are still a number of practical limitations unique to medical imaging. Particularly supervised DL methods (which are the most common) for focal pathology segmentation rely heavily on subjective annotations, which can cause a number of complications. Subjective annotations can lead to biased DL models that not only fail to generalize, but also fail to accommodate the specific needs and demands of different medical segmentation goals.

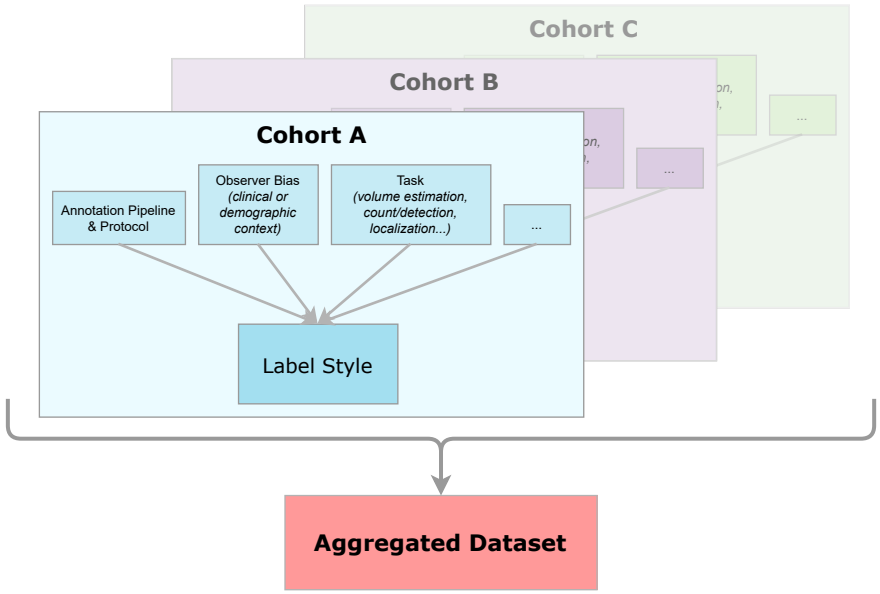
Absolute "ground truth" annotations are often not attainable in pathology segmentation tasks due to the limits of certain acquisition sequences, partial volume effects, as well as the ambiguity in the border of focal pathologies (ie lesions or tumors). Therefore, these annotations are subject to high *inter-rater* variability, even when annotated by the most skilled expert raters. Establishing a fixed annotation process can help increase inter-rater

reliability, but also introduces biases that are embedded into the annotation process itself. Biases can also arise from semi-manual labelling, a process by which automated methods are used to generate a preliminary label that is then corrected by a human rater. Therefore the automated algorithm used in this process can introduce its own biases. The goal of the study (e.g. diagnosis, counting lesions, volumetric measurements), or the instructions provided to raters, and many other things can contribute to biases in the annotations. Since performance metrics are computed with respect to these variable types of "ground truth" annotations, model performance must be interpreted with some care. As annotation processes vary between datasets, attempts at generalizing to new datasets may be misguided as researchers may be performing unfair comparisons between incompatible ground truths, therefore sullying the evaluation results. Specifically, it is not fair to compare the predictions of a model that has learned definition of "truth" to the labels from a dataset that has a differently defined "truth". This will result in bad performance metrics, which may be interpreted as a model having poor generalizability, when this may not be the case. With this in view, the whole concept of generalization must be rethought, particularly when generalization performance is measured on hold-out datasets which may be subject to completely different annotation biases.

Researchers often hope that algorithms trained with a large, diverse dataset will generalize well to new data of the same task. However, ill defined boundaries due to the heterogeneity of the pathology makes lesion segmentation highly prone to variable annotations. This proves especially detrimental in efforts to deploy pathology image segmentation models to real world settings. Consider an example where Hospital A is focused on lesion counts in multiple sclerosis (MS) patients, and where Hospital B is measuring lesion volumes in MS patients as depicted in Figure 1.1a. Both tasks involve lesion segmentation, but the resulting "ground-truth" labels in each case are quite different. Hospital A's style results in more, but smaller lesions, while Hospital B's style results in larger and fewer lesions. While differences of this kind are often chalked up to inter-rater variability, this view ignores other, often confounding, factors that may bias the annota-



(a) Example case demonstrating how changing the labeling task per cohort can affect the final labels.



(b) Example depicting several factors contributing to annotation shifts and label styles in aggregated datasets.

Figure 1.1: Illustration of sources and problems caused by label biases in aggregated datasets.

tion process. This example demonstrates the issue of *annotation styles* in focal pathology segmentation which, due to the lack of absolute "ground truth", are subject to not just inter-rater variability, but to biases embedded into the annotation process itself. As a

result, model performance on a hold-out dataset with a different "ground truth" annotation style will inevitably see a performance drop, not necessarily due to a distribution shift in the image-space (scanner, image protocol, etc.), but due to a distribution shift in the label-space. Although there have been a number of successful papers that overcome a performance drop due to image-space distribution shifts [7, 25, 38, 78, 92], overcoming a performance drop due to different "ground truth" annotation styles remains an open problem.

Several other researchers have noted the problem of annotation styles in large datasets and the impact on generalization [12,66]. The example provided previously demonstrates one type of annotation style difference and its impact on generalization; however, in reality, many different factors contribute towards annotation styles. For most medical segmentation tasks, much of the labelling is done semi-manually, where differences in the software or even software version can have a significant influence on the final label. The degree of correction done to the automatic labels may also vary from dataset to dataset. Furthermore, depending on the specific aim of the clinical task, the labelling process may differ, even if the end label is still a segmentation mask. For example, certain clinical trials may have instructions to generate labels in specific ways to obtain markers for evaluating treatment effect, as illustrated in Figure 1.1a. Additionally, some contexts, such as clinical trials, prioritize consistency in labelling, while others may consider each patient sample independently and are not focused on consistency. For instance, while Hospital A might elect to use a fully manual annotation process for detection, Hospital B may elect to use semi-manual annotations given that inter-rater variability would otherwise obfuscate typical changes in total lesion volume across scans thus not obtaining the consistency required for the task. Even within a constant labelling protocol, previous works have identified annotation shifts caused by inter-rater variability [11, 36, 80, 94]. In cases where both the raters and the labelling protocol are consistent, there still exists the issue of observer bias, wherein the rater uses their own medical knowledge about a patient or patient cohort to inform their labelling decisions. In these situations, raters can also

use other information in the image to influence their annotations based on clinical priorities or significance. In many cases, these biases cannot be isolated as they compound throughout the labelling process, therefore making them particularly difficult to model.

The resultant annotation styles can thus affect perceived performance especially in generalization challenges. Bias-invariant or domain-invariant methods are becoming more popular in generalization tasks, especially for image-based distribution [29, 37, 46, 86]. However, the impact of annotation styles or biases in the label space can still prove a formidable challenge for these methods. Domain-invariant approaches are incredibly useful in learning robust feature extraction from the images, but since labels are dependent on the circumstances of their generation, the final label used to evaluate generalization is still inherently domain-dependent. Due to this complication, being completely invariant to the source domain can lose information regarding annotation style that effects the final results [8, 97]. Utilising information relevant to annotation style with automated algorithms can therefore help researchers perform informed evaluations and obtain the labels desired for their task. Consider a scenario with Hospital A and Hospital B again. What if these organizations attempted to pool their data together in order to obtain a large and varied dataset to train an automated algorithm? Being independent of the source of the dataset may result labels that do not reflect each institution's respective goals. Being able to aggregate the two datasets and learn from the additional data while still generating the desired annotation style for both hospitals requires meta-information regarding the annotation process. If another institution, Hospital C, had the same annotation process as Hospital A, that would also be valuable information. This key piece of information could notify the DL practitioner that the algorithm can learn the annotation style for Hospital A by considering both Hospital A and Hospital C as one style. The overall benefit of considering auxiliary information relevant to the annotation style is that the model can more strategically leverage data from multiple sources, while accommodating each source's unique requirements and performing a fair evaluation.

In this thesis, we explore annotation styles and their influence on DL networks in the context of pathological segmentation. We also present a simple, in-line method to account for and to model annotation styles in aggregated medical datasets. We propose to modify a U-Net architecture by adding Conditional Instance Normalization (CIN) [93] to learn different annotation styles across several different multi-center, multi-source, MS clinical trial datasets. With CIN, we provide the input images alongside auxiliary information to condition on, which allows us to learn the relevant styles that are not clearly predictable from the image. We then have a singular model capable of producing many possible outputs for a given input image in various annotation styles. This functionality is especially useful in MS studies. Due to the nature of MS lesions, they have proven to be a challenging task to segment what with ambiguity in lesion borders as well as intermediately diseased tissue, and scarring. By providing multiple possible segmentation predictions, our method can be tailored to specific needs of different healthcare centres, as well as provide additional information to healthcare practitioners. An overview of the general method is shown in Figure 1.2.

With our propose framework, we perform an extensive study of annotation styles and their impact on generalizability, and further examine and identify relationships between annotation styles. To do so, we compare our conditioned models to *Naive-Pooled* models, which aggregate datasets and provide no context for which dataset an input image came from (no CIN), and *Single-Trial* models, which are trained on only one dataset/trial. Lastly, we propose a method to fine-tune an existing network to new annotation styles for sample-efficient continual learning strategies.

1.1 Multiple Sclerosis

MS is a progressive and inflammatory disease that results in demyelination in the central nervous system (CNS) [16, 59, 72]. Symptoms of MS vary greatly, from weak limbs, numbness, blurred vision, dizziness, fatigue, and poor motor control. Canada has one of

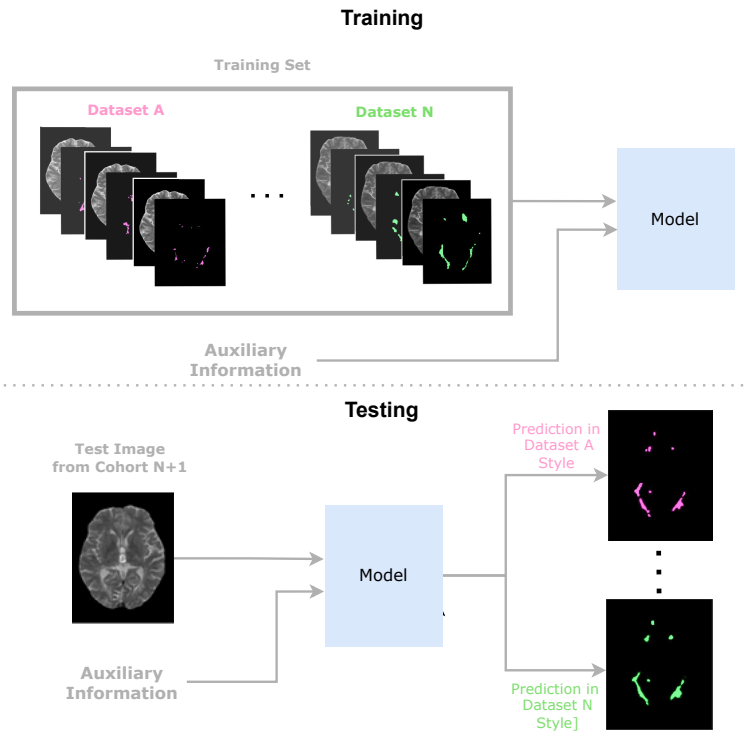


Figure 1.2: System overview. (Top) Training module: Training on multiple cohorts with auxiliary cohort information to learn the associated bias for each cohort. (Bottom) Testing module: Auxiliary cohort information used to generate multiple lesion segmentation maps, each with a different label style, for the input test image.

the highest rates of MS in the world, with 1 in every 400 people suffering from MS. Symptoms can follow a relapse and remission pattern or a progressive pattern, or a type of intermediate combination [16]. Globally, females are twice as likely to get MS than males, and in Canada, females make up over 75% of cases. MS affects a relatively younger population, with the global average age of diagnosis being 32 years. Its the most frequently occurring demyelinating disorder among young adults and is a leading cause for non-traumatic neurological disability in young adults [16, 72]. The timing of onset results in MS affecting people during important career- and life-building years [55, 59]. Estimates in 2020 predict an average unemployment rate of 60% among Canadians with MS [55]. A recent epidemiological model created in 2017 predicted that MS cases will rise from 4051 cases per 100 000 in 2011 to 4794 per 100 000 in 2031. They also predict that total

annual healthcare sector costs will reach $ 2.0 billion by 2031. These factors make MS a very pressing issue for healthcare in Canada [59].

One way to improve healthcare and quality of life for people living with MS is early detection and intervention. Early diagnosis and treatment can potentially reduce relapses and disability [72]. The McDonald criteria was originally published in 2001, and has since been updated every few years, with the latest being released in 2017. Two key biomarkeres for MS include 1) Dissemination in Space (DIS) which is characterized by inflammation/lesions occurring in different regions of the CNS, and 2) Dissemination in Time (DIT) which is characterized by recurring inflammation in the CNS [72, 83]. Both DIT and DIS need to be present to diagnose MS, and both are detectable through magnetic resonance imaging (MRI). As of the 2010 update to the McDonald Criteria, MS can be diagnosed from one baseline MRI scan with asymptomatic contrast enhancing lesions. These lesions appear as T2 hyperintensities, or can be enhanced using a contrast agent, such as Gadolinium [72]. The borders of these MS lesions are an intensity gradient, and as a result are subject to ambiguity [14]. Therefore, clinical MRI and tools for consistently and reliably identifying lesions plays an important role in diagnosis and disease monitoring, as well as treatment evaluation [72].

MS has several different disease courses, also termed disease pheotypes. The relapsing form of MS is commonly called Relapsing Remitting MS (RRMS), and the two progressive types are Primary Progressive (PPMS) and Secondary Progressive (SPMS). Lastly, there is an intermediate form termed Progressive Relapsing MS (PRMS), although this diagnosis is not commonly given [52]. RRMS patients suffer from recurrent attacks of neurological dysfunction, usually followed by at least partial recovery. Frequency and intensity of attacks can vary, and although partial recovery is observed between attacks, many relapses may never completely revert and patients incur disability [28]. SPMS may eventually develop in patients with RRMS. SPMS is characterized by progressive clinical disability in the absence of relapses. The transition from RRMS to SPMS is gradual and not clearly defined. PPMS is characterized progression in the absence of relapses from the onset of

the disease, i.e., without a relapsing remitting stage. SPMS and PPMS subjects are about 10 years older than RRMS patients on average. SPMS patients have larger lesion volumes than RRMS and PPMS patients [52]. Identification of progression in MS is done retrospectively, but as previously mentioned, the exact lesion segmentation mask is difficult to establish. Often, progression is identified by slowly-expanding or enhancing lesions from MRI [17].

Several new treatments have become available in the last 20 years, with new therapies being consistently investigated and developed. Current treatments are more common for relapsing forms of MS. Treatments are primarily aimed at reducing and preventing attacks, and mitigating debilitating symptoms [24, 56]. However, there is an increasing number of MS treatments being approved for progressive forms of MS [17]. Advancements in recent treatments have been able to improve quality of life for people with MS and reduce long term permanent disability. There are disease-modifying therapies which modify the function of the immune system and reduce inflammatory activity, thus reducing rate of relapses and reducing accumulation of MRI lesions.

Diagnosis, progression tracking, treatment planning, and treatment development all rely heavily on MRI and accurate lesion identification. Full segmentation of T2 lesions can provide accurate volume assessments, counts, and identification new and enlarging lesions for various clinical aims [52]. Depending on the clinical aim of a given MS dataset, the annotations can look different, as illustrated in the Hospital A and B scenario. Furthermore, the ambiguity of MS lesions leads to different annotation styles depending on the specific definition or guidelines used at a given center or in a given dataset. As a result, automated methods for MS T2 lesion segmentation need to be able to accommodate and handle a variety different annotation styles. These automated methods also need to be able to adapt to new, changing requirements for MS lesion identification as knowledge about the disease evolves. As each MS clinical trial may be looking for different treatment effects, for example, reducing new and enlarging T2 lesion counts, or reducing increased

T2 lesion volume, automated methods also need to be able to produce the annotation styles necessary for these analyses.

1.2 Contributions of Thesis

In this thesis, we will demonstrate the existence as well as the impact of annotation styles across our MS clinical trial datasets. We also implement CIN in a novel framework to account for said annotation biases, enabling efficient use of multiple clinical trials while also allowing for more fair generalizability assessments. With our method, clinics will be able to leverage more data while obtaining predictions tailored to their specific needs. Through a series of experiments, we will show and explore the following:

1. **Exposing the effect of annotation styles on deep learning segmentation algorithms and generalizability.** "Ground-truth" annotations are required for all automated supervised learning methods. The process for obtaining these annotated datasets often requires generation of labelling guidelines or rules, the use of semi-automated software, and the input of human rates. These factors can contribute to the development of annotation styles. Often, DL methods do not acknowledge or address annotation styles. This can lead to unfair evaluation of model performance, especially in evaluating the generalizability of a model to new datasets (and consequently, new annotation styles). The presented results show that differences in annotation styles still persist even when disease, disease phenotype, and image preprocessing pipeline are all kept consistent across different datasets. Furthermore, this thesis demonstrates the problem created by simplifying assumptions that ignore annotation styles in generalizability assessments. This thesis proposes a novel use of CIN for modelling such annotation styles across aggregated datasets, effectively leveraging more available data and allowing for fair generalizability evaluations. This thesis shows that annotation syles can be an especially prevalent problem

in MS segmentation datasets given the unique and challenging ambiguities of the lesions.

2. **Presenting a new method for identifying similar annotation styles across different datasets for use in strategic dataset aggregation.** Many medical datasets are very small due to costly acquisition, privacy concerns, as well as small sample sizes for some rarer pathologies. This can make learning separate annotation styles for every dataset difficult when some datasets are incredibly small. To combat this, this thesis proposes a new post-hoc analysis method based on our conditioning framework to identify similar annotation styles across different datasets. This thesis then shows that by pooling datasets with similar annotation styles and treating them as one dataset in a new model, we can obtain competitive performance with the desired annotation styles. This makes more efficient use of each dataset by providing the model with more samples for each annotation style, allowing it to learn said annotation style with more accuracy.

3. **Fine-tuning to annotation styles with few labelled samples for continual learning.** As new dataset are collected from continuing advancing research, new annotation styles are inevitably going to emerge. Research can also change the definitions or interpretations of medical pathologies, which can also contribute to necessary changes in annotation styles. Algorithms implemented in real-world applications like clinics or medical research facilities need to be able to produce the desired annotation style, even it if is an unseen one. To fulfill this need, this thesis proposes a fine-tuning strategy to quickly adapt an existing model to a new annotation styles with only a few labelled samples. This fine-tuning strategy will reduce the need for large manually or semi-manually annotated datasets, thus saving time and money for the healthcare sector while still meeting the individual needs for separate studies.

1.3 Thesis Overview

This thesis provides a useful guide for researchers on annotation styles and their potentially damaging effects on automated DL segmentation methods, and how to deal with them. We present a generalized conditioning framework and a fine-tuning approach to easily account for shifts in annotation styles. Our proposed methods are evaluated on our real multi-center, multi-source, MS clinical trial datasets.

Chapter 2 covers relevant academic background on DL, medical segmentation, and related works. This chapter will cover background information on all components of the DL networks implemented as well as the corresponding training processes used in this thesis. This chapter will also go into detail on the common "ground truth" generation processes. Next, a literature review on DL in medical segmentation and MS segmentation is presented. Related literature in the domain adaption field that focuses on adapting to or accounting for other types of biases that occur in medical datasets is also covered. Lastly, related fine-tuning adaptation research is summarized.

Chapter 3 explains the details of the various methods used in this thesis. This chapter covers Conditional Instance Normalization (CIN), the proposed method for modelling annotation styles in medical segmentation DL networks. Here, I explain how CIN is incorporated into an nnU-Net for annotation style adaptation. The Naive-Pooled model and Single-Trial models used in this thesis are also defined in this chapter. This chapter presents a new post-hoc parameter analysis method using cosine similarity to study the relationships between learned annotation styles. Finally, details are provided for the fine-tuning approaches proposed in this thesis to adapt existing networks to new annotation styles using only 5 labelled samples.

Chapter 4 This chapter covers implementation and experimental specifics. First, it describes in detail the datasets used, including their trial identities, years collected, and associated disease phenotypes. The experimental approach using the Naive-Pooled and Single-Trial models as baselines and as analysis tools is also described. Lastly, imple-

mentation details and hyperparameter optimization approaches are also described for reproducability purposes.

Chapter 5 investigates the influence that annotation style has on the generalizability of DL algorithms. This chapter uses our CIN-based model and condition on the trial identity (termed *Trial-Conditioned*), and compare to Naive-Pooled and Single-Trial models in a mixed-phenotype dataset. Here, the outcomes produced by the Trial-Conditioned model are used to demonstrate that the differences in annotation style can and should be modelled in order to perform fair evaluations of DL networks. The results show that Single-Trial models fail to generalize to other annotation styles, showing a performance degradation of up to 10% relative to models that have learned the target annotation style. Similarly, the findings demonstrate that Naive-Pooled models can fall behind Trial-Conditioned models by $\tilde{2}$-3% while also failing to properly learn each trials unique annotation style. Furthermore, the results demonstrate that failure to use the correct annotation style in a Trial-Conditioned model results in performance degradation of over 10%. This chapter continues to investigate this impact of annotation styles on phenotype-consistent datasets, with similar Naive-Pooled, Single-Trial, and Trial-Conditioned models. The chapter then shows the existence of different annotation styles even across these phenotype-consistent datasets. The results from these experiments all confirm the existence of different annotation styles, even across datasets with the same phenotype and identical image preprocessing pipelines. These results allow us to further understand annotation styles and the potential impact that disease phenotype has on resulting annotations, whether it be by observer bias or by significant differences between disease expression. Lastly, with a mixed-phenotype aggregated dataset, we compare a *Phenotype Conditioned* to a Trial-Conditioned model to determine the relative impact of both individual trial collection and disease phenotype on annotation style.

Chapter 6 presents the results of the proposed cosine similarity method to discovery similarities and subgroups between annotation styles. The utility of the identified subgroups is then demonstrated using another series of pooling and conditioning exper-

iments. This chapter shows that we can successfully and efficiently pool trials using the automatically detected annotation subgroups while maintaining performance, thus validating the grouping method as well as demonstrating its utility in data-scarce situations. By conditioning on the group identity (*Group-Conditioned*), we obtain competitive performance with the Trial-Conditioned model. This Group-Conditioned model is useful in data-scarce situations where each individual dataset may not have enough samples to effectively learn the style.

Chapter 7 presents the results of the proposed fine-tuning methods. With these fine-tuning approaches, this thesis is able to adapt a pre-trained model to new annotation styles with only 5 labelled samples.A detailed analysis of the different proposed methods and their impacts on the learned annotation style is shown. This chapter applies the fine-tuning method to both a Naive-Pooled model and Trial-Conditioned model to demonstrate the versatility of the proposed fine-tuning approach. This chapter details the benefits and uses of our proposed fine-tuning approach for life-long learning and clinical implementation of DL models.

Chapter 8 concludes this thesis by summarizing the key findings and arguments presented in Chapters 5 through 7.

Chapter 2

Background and Related Works

This chapter first provides a foundational background in DL mechanics. We present the basics needed to understand the architecture and training of DL networks. We then provide information on DL and image segmentation, followed by DL in medical segmentation. This section describes various limitations that effect DL in medical segmentation, particularly in relation to the ground-truth annotations. We discuss some related works that also aim to address other limitations in DL medical networks caused by various inconsistencies in datasets. As annotation styles are not commonly studied, we present the closest related literature on topics such as inter-rater variability, domain adaptation, and style adaptation in order to provide context of the field and how this thesis fits into the scope of various issues more commonly addressed in medical segmentation with DL.

2.1 Deep Learning

Deep learning (DL) is a sub-field of machine learning focused on multi-level representational learning techniques. Generally, conventional machine learning methods require feature engineering or other processing techniques to work adequately, but DL techniques are able to work with more complex data forms with relatively little preprocessing. DL methods have drastically improved state-of-the-art performance in many tasks such as

natural language processing, genomics, and especially computer vision. There are two main types of DL (and machine learning): supervised and unsupervised. Supervised learning is the most common type of DL and it requires both the input data and the correct "answer" to the task for learning [48]. This correct "answer" used in both training and evaluation metrics is termed the ground-truth.

2.1.1 Convolutional Neural Networks

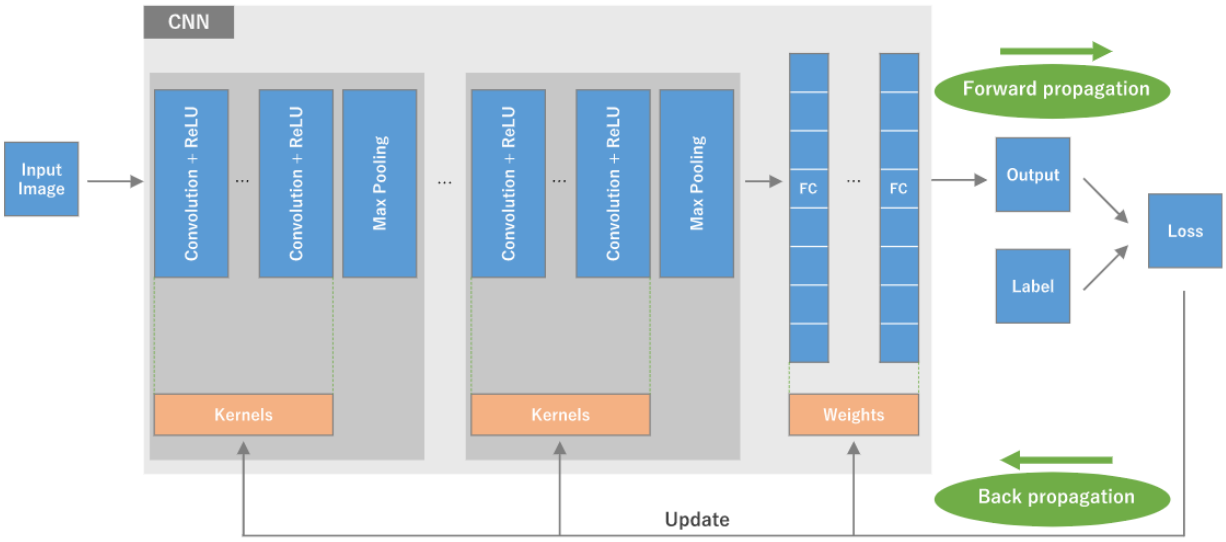


Figure 2.1: Simplified general convolutional neural network consisting of the common building block layers of many neural networks. Figure courtesy of [100].

The most common and most established type of DL network is the Convolutional Neural Network [100]. CNNs are designed primarily for imaging data (or grid-pattern data), and consist of layers of mathematical processes to extract both high- and low-level features from the input. The input can be one individual sample, or it can be a *batch* of samples fed to the network simultaneously. In cases where the batch size is greater than 1, the network can output the predictions of the whole batch at once. CNNs typically consist of convolutional layers, activation functions, pooling layers, and fully connected layers. While convolutional and pooling layers are used for feature extraction, fully connected layers are used at the end of the network to transform the outputs into arrays for

classification tasks. An overview of a simple classification CNN with the basic layers and functions is shown in Figure 2.1.

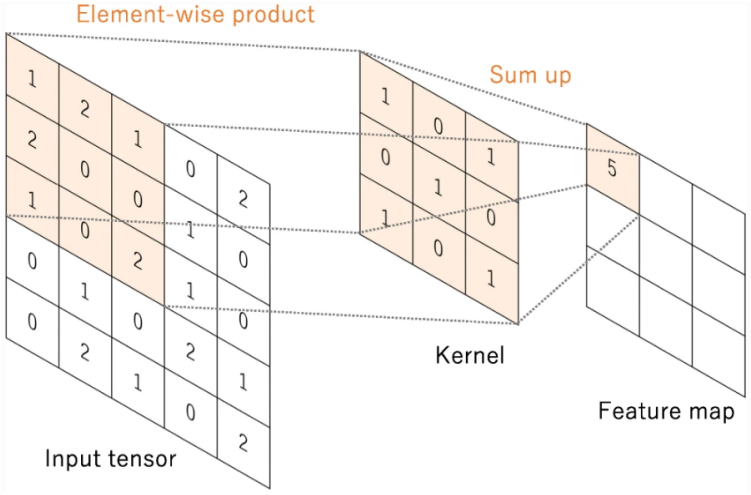


Figure 2.2: Convolution operation example with a kernel size of 3x3, no padding, and a stride = 1. Figure courtesy of [100].

The foundational block of the CNN is the convolutional block. The convolutional block is typically convoltion operator paired with an activation function and often a normalization function. The convolutional operation is shown in Figure 2.2. In convolution, a *kernel*, which refers to a matrix of learned parameters, is multiplied element-wise by the input tensor and the resultant product is then summed to produce a feature map. This kernel is "slid" across the input tensor by a value of *stride* to complete the feature map [100]. This feature map output from the convolution is then fed to a non-linear activation function in order to give CNNs capacity to model non-linearities. Common activation functions include the ReLU, leaky ReLu, Sigmoid, Tanh, and Softmax [77]. The normalization function can either follow or precede the activation function, depending on the specific design of the CNN. There several types of normalization layers, but the two most common types are batch and instance normalization. Batch normalization (BN) normalizes the input to a zero mean and a standard deviation of 1 according to the statistics of the entire batch. Instance normalization (IN) normalizes the input according to the statistics only of the input.

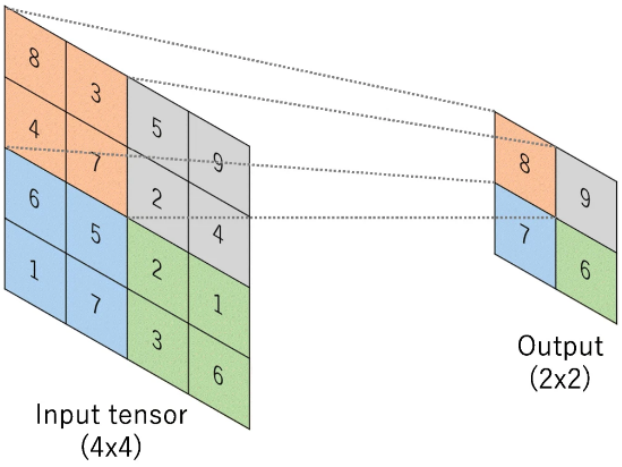


Figure 2.3: Maxpool operation with a filter size of 2 and a stride = 2. Figure courtesy of [100].

A pooling layer is designed for the down-sampling of feature maps. A Maxpool layer is shown in Figure 2.3. Max pooling takes the maximum value of the values within the filter area and uses it in the output matrix. Similarly, average pooling takes the average of the values in the filter area and uses it in the output matrix.

2.1.2 Training Convolutional Neural Networks

Training a CNN is the process of iterative updating the parameters of the network to reduce the difference between the predicted output from the CNN and the ground-truth. This process is done with what is called the training set. In DL, a given dataset is often split into a training, validation, and testing set. The training set is used for training, and the validation set is used as an approximation of how the model performs on unseen data between *epochs* of the iterative training process. The test set is reserved for the final trained model to assess its real performance on unseen data.

During training, the difference between the predicted output from the CNN and the ground-truth is calculated using a *loss function*. Common types of loss function include Cross Entropy loss, Tversky loss, and Focal loss [32]. All loss functions must be differentiable as the gradient back propagation method is used with the loss function to update

the network parameters. An optimization function is used to calculate how to change the parameters of the network given the loss. The optimization function uses the derivative of the loss with respect to the parameters of the network with an update rule to iteratively update the parameters. The *learning rate* of an optimization function is the value by which the parameters will change. There are several options for optimization functions including stochastic gradient descent, momentum, RMSProp, and Adam [10].

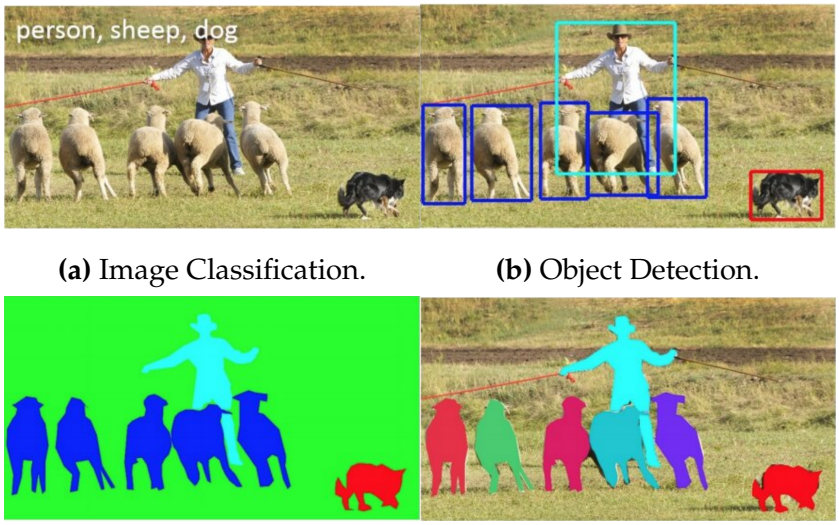
Training continues until the performance on the validation set plateaus. In some cases, the performance on the validation set may begin to decline even though performance on the training set is still increasing over epochs. This scenario occurs when the model *overfits* on the training set and therefore its performance on new data degrades. Overfitting is often combated using dropout or data augmentation. Dropout is when units of the neural network are randomly set to 0 to prevent over-adapting to the training set. Units can be channels or individual activations [85]. Another strategy to prevent overfitting is data augmentation, the process of randomly changing the input data during training. Data augmentation not only induces realistic variations in the data for improved generalizability, but it also artificially increases the dataset. In computer vision, data augmentation operations include image flipping, rotations, contrast operations, or colour channel changes [96].

Hyperparameter Tuning

A hyperparameter of a DL network is different from a network parameter in that it is not learned during training, and it is pre-set by the researcher. Kernels, and other weights or biases within the network layers are learned during the training process described in the previous section. Other aspects of training are selected by the researcher, such as: the batch size, loss function, optimizer, learning rate, the number of training epochs, dropout probability, and the data augmentation policy used. Furthermore, many design decisions of the network are also considered hyperparameters, such as the number of layers in the

network, and the kernel sizes, strides, and padding. These hyperparameter decisions are made on the basis of validation performance [100].

2.1.3 Deep Learning for Image Segmentation



(a) Image Classification. **(b)** Object Detection.

(c) Semantic Segmentation. **(d)** Instance Segmentation.

Figure 2.4: An example demonstrating difference between image classification, object detection, semantic segmentation and instance segmentation. Figure courtesy of [53].

Image segmentation is an image labelling process that partitions an image into segments. It classifies each individual pixel to generate a segmentation mask. It is an important aspect in visual understanding and is required for many different computer vision engineering tasks from medical imaging, to augmented reality [53]. Image segmentation can fall under several categories: binary segmentation, semantic segmentation, or instance segmentation. Binary segmentation is the simplest, where there is only one class or object to be segmented, and the rest of the image is labelled as background. Semantic segmentation is essentially a multi-class extension of binary segmentation, and instance segmentation involves segmenting different classes of objects, as well as different individuals of such objects. Figure 2.4 shows classical computer vision examples of semantic and instance segmentation in comparison to the other popular image labelling methods

including object detection and image classification. CNNs have become particularly popular for image segmentation tasks due to their grid-oriented design and the lack of a feature engineering required. Although the first CNN for image analysis, AlexNet [45], was geared towards classification, CNNs quickly became popularized for segmentation. One of the earliest CNNs for image segmentation was adapted from an object detection architecture in 2014, named R-CNN [21]. Since then, CNNs have become a staple in image segmentation.

2.2 Evaluation Metrics

Any binary tasks are all evaluated on the basis of a binary confusion matrix. The binary confusion matrix describes the basic prediction evaluation outcomes and is shown in Figure 2.5. For binary segmentation, these outcomes are on a voxel-by-voxel (or pixel-by-pixel for 2D) basis and as such, other metrics are designed to summarize the performance over an image, and are averaged over a testing or evaluation dataset.

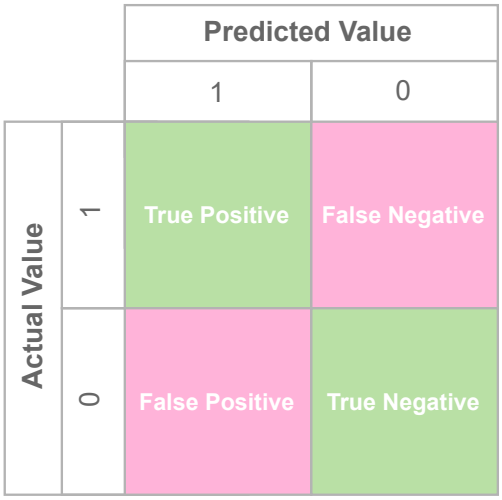


Figure 2.5: Binary confusion matrix defining True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

DICE score, also called F1 score, is commonly used to evaluate segmentation performance in medical imaging tasks and quantifies the *overlap* between the predicted label

mask and the true label mask [87]. The formula for DICE score calculation per image is shown in 2.1.

$$Dice = \frac{2TP}{2TP + FN + FP} \tag{2.1}$$

Recall, also called True Positive Rate or Sensitivity, is the rate of correctly predicted positives out of all the total positives in the true mask [87]. This metric is especially useful in cases where detecting all possible positives is very important; however, it does not provide any information on FPs. As a result, judgement by this metric alone can lead to selection of algorithms that over-segment. For this reason, this metric is often paired with Precision.

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

Precision, or Positive Predictive Value or Specificity, is the rate of correctly predicted negatives. Since Precision penalizes FPs, it is beneficial to use in cases with a high class imbalance where the majority class is background/negative [70].

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

The above metrics all rely on the selection of a *threshold* to binarize the predicted result. Selection of threshold can heavily influence the performance metrics, so in order to provide a metric that does not rely so heavily on selected threshold, we also use Precision Recall Area Under Curve (PR-AUC).

A precision recall curve is the plot of precision vs recall at different binarization thresholds, and so it follows that PR-AUC is the area under this curve. We obtain PR-AUC for

a very wide range of thresholds, therefore allowing us to evaluate performance without high dependence on threshold selection. Often Reciever Operating Curve (ROC) AUC is used, but Precision Recall curves are more useful for imbalanced datasets [15]. Since this thesis focuses on lesion segmentation, which is highly imbalanced between lesion class and background class, we opt to use PR-AUC.

2.2.1 Generalization Evaluation

Evaluating the generalizability of a network is an important step in the design process, as a network needs to be applicable to a wide range of data for it to be useful in real-world scenarios. Generalizability of an algorithm is often evaluated based on the performance metrics achieved on the held out test set. Since the held-out test set comes from the same dataset, it may not provide a true estimate of an algorithm's ability to generalize to data collected from different environments or with different methods. Furthermore, DL models are so high capacity that they have been shown to be able to fully memorize random labelling of training data. Recent research suggests that even with use of heavy regularization, models can still memorize random patterns [103]. As a result, common generalization evaluation results need to be considered with care.

2.3 Deep Learning in Medical Image Segmentation

For medical applications, medical image segmentation is the classification of voxels from medical imaging modalities. These regions or structures can include anatomy or pathological structures such as lesions or tumors. Segmentation of structures can allow for calculation of important metrics such as total volumes, lengths, counts, and more. Segmentations are useful for a wide range of medical applications including but not limited to diagnosis, anatomy or population studies, diseased tissue localization, treatment planning and evaluation, and computer aided surgery [64]. Segmentation can be performed on a variety of imaging modalities depending on the application from retinal photogra-

Figure 2.6: Top row shows the original images, and the corresponding images on the bottom row is the segmentation mask. From left to right is skin cancer, lung, retinal vessels, and prostate. Figure courtesy of [33].

phy, positron emission tomography, X-Ray, ultrasound, MRI, and computed tomography (CT). Some examples of medical segmentation are shown in Figure 2.6. Some segmentation tasks are more ambiguous or challenging than others. Particularly, consider the skin cancer image shown in Figure 2.6 where the border is irregular and the contrast is poor. Exact delineation of the border may be more challenging and subjective compared to the border of the lungs shown in the 2nd column.

Although previously, medical image segmentation was done with methods like atlas-based methods, clustering algorithms, preprocessing and thresholding methods, and region growing approaches [76], deep learning networks have become increasingly popular in medical segmentation tasks. Particularly after the introduction of the widely successful U-Net CNN architecture [68], the field of medical computer vision developed a lot of interest in deep learning networks. The U-Net, first proposed in 2015, follows an encoder-decoder structure as shown in Figure 2.7. The U-Net, alongside with small variations of the U-Net, have been used extensively in a variety of medical segmentation tasks. U-Nets have been used for anatomical tasks such as ultrasound fetal femur measurements [98], eye tissue segmentation from RGB images [22], ventricle segmentation from echocardiography [54], and many more. U-Nets have also been applied quite successful in many

pathology segmentation tasks including breast cancer calcification [84], brain tumors [1], liver cancers [5], and MS. Despite the many variations proposed of the U-Net and the many other new networks proposed since its introduction to the field, recent research has demonstrated that well-trained traditional U-Net is still state-of-the-art. Research done by Isensee *et al* [30] has shown that a well engineering, well trained U-Net with only some minor modifications (such as use of Instance Normalization and slightly different activation functions) is incredibly difficult to beat. They have won or placed very high in many segmentation competitions, and their findings have been validated by other researchers [60].

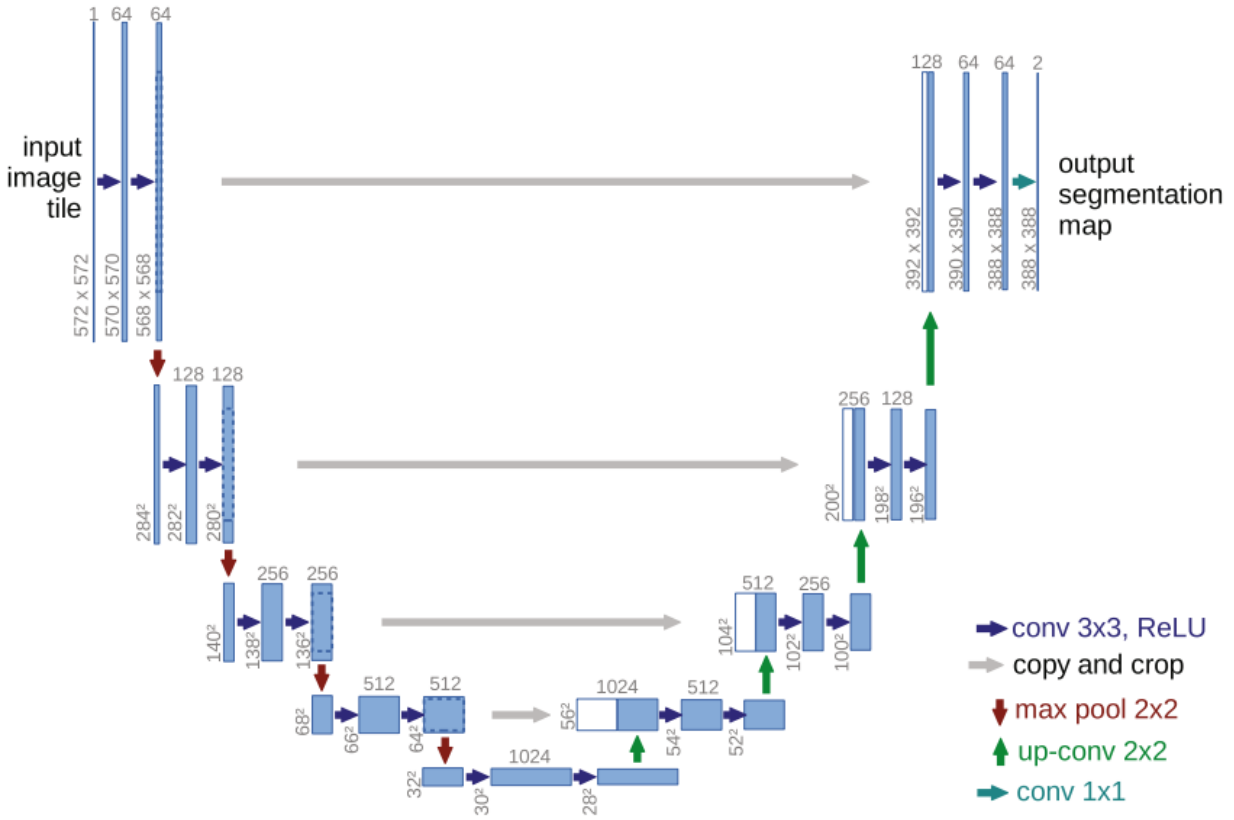


Figure 2.7: Original 2D U-Net architecture with 32x32 pixels in lowest resolution. Each blue box represents a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations as defined in the legend. Figure courtesy of [68]

.

2.3.1 Challenges in Generating Medical Ground Truth Annotations

Training, validation, and generalizability assessments of supervised DL models all rely heavily on ground-truth annotations. In natural imaging tasks, ground-truth annotations are relatively objective, as long as imaging quality is adequate. As shown in Figure 2.4, the labels for natural images are relatively intuitive, and the average person would be able to produce accurate ground-truth annotations with ease. In many cases as well, the variations in natural imaging ground-truth labels are likely caused by simple errors [75]. In medical segmentation, ground-truths are not as objective. Understanding medical images often requires more expertise, and many medical imaging modalities represent different tissue types with subtle differences in intensity. As a result, medical image "ground-truth" annotations can become subjective, and thus effected by many different factors, resulting in the aforementioned *annotation styles*.

Several factors can contribute to annotation styles in ground-truths. Researchers have identified inter-rater bias or variability as a fairly common occurrence in medical image segmentation tasks [26, 35, 36, 94, 99]. [18] also found that gaps between expert intuition can affect annotations, stating that some raters simply relied on their intuition which could not be explained. Another reason factor is due to the subjectivity of "ground truth" in medical pathology [12]. Although related to inter-rater bias, subjectivity is a particular problem in pathological applications where a true "ground truth" is not necessarily attainable. Observer bias, which is integrated in rater bias, is also a contributor which is often very implicit and not explainable. A study conducted by [27] found observer bias caused significant difference in labels generated by raters that were fully blind compared to non-blinded raters in a clinical trial dataset. Blinding is often thought of as useful for eliminating confounding factors; however, in the case of medical segmentation, [18] remarked that some annotators that are blind to medical history may produce potentially "unacceptable" annotations compared to raters with access to full medical records. To, or to not provide medical records to the raters then further contributes to differences in possible ground-truth annotations.

Furthermore, the label generation process itself can be a major source of bias or contributor to annotation style [79]. [18] noted that the guidelines provided to raters had an impact on quality and annotation outcome. [95] also found that differences in labelling protocols can degrade algorithm performance, as well as limit to what extent performance can be validated. [63] recommend that labelling rules and other data generation procedures be made available to users due to their impact on testing results. [12] further discuss the impact of the use of semi-automated labelling software in the label generation process and resulting labels. Even when dataset generation processes are kept as consistent as possible, due to the influence of human annotators and software differences, researchers still find differences in annotation styles of datasets. This kind of problem with annotator differences is even noted in natural imaging tasks, despite the significant relative increase in objectivity [67].

2.3.2 Multiple Sclerosis Lesion Segmentation

There are several MS segmentation tasks including Gadolinium lesion segmentation, T2 lesion segmentation, and new and enlarging T2 lesion segmentation. [13] used 3D U-Net for segmentation of Gadolinium lesions, and [71, 74] used U-Net based methods for segmenting and detecting new T2w lesions. [4] performed MS T2 lesion segmentation with a modified ResNet-based segmentation network, and [19] used a multi-class U-Net for lesion and brain tissue segmentation. Although there are multitudes of ways to segment MS patient images, Gadolinium lesions are much smaller and the labels are more prone to noise [13], and new or enhancing T2 lesions require two or more labelled images from different time points, thus making them less ideal candidates for analysis of annotation style problems compared to T2 segmentation.

Multiple Sclerosis-Specific Annotation Challenges

As previously discussed, medical segmentation tasks have unique difficulties affecting ground-truth annotation that are not common in natural image analysis. MS ground-

truth annotations can be affected by all aforementioned factors; however, MS presents even more unique challenges as a complex neurological pathology. Specifically, [2] have noted that lesion segmentation in MS, especially new and enlarging lesion segmentation, can be highly disagreed upon by raters. Furthermore, the border of MS lesions represents a continuous (rather than discrete) transition in voxel intensity from the surrounding normal tissue to the lesion. The ambiguity of MS lesion borders is further amplified by the existence of an intermediate, pre-lesional abnormality called diffusely abnormal white matter (DAWM) [14]. Because there is no universally accepted definition for DAWM, some raters might include more DAWM within their lesion masks than others, leading to arbitrarily different annotation styles. Even within a lesion, there are inherent inconsistencies of the intensity profile. Depending on display contrast when a rater is viewing an MRI for manual segmentation, the lesion size may appear slightly larger or smaller than for another rater who might be viewing under different display conditions. These factors all contribute to the additional unique challenges of MS ground-truth annotations, thus exacerbating the potential for annotation style-related problems.

2.3.3 Limitations

Although DL methods have shown great success in medical segmentation tasks, there still exist many practical and implementation based concerns. Many of these concerns revolve around the data. Especially in the medical field, data is hard to obtain, and supervised DL models require very large annotated datasets. The aforementioned variations in the so-called "ground-truths" of these annotated datasets are a newer forthcoming issue that is still under investigated in many contexts; however, other concerns for reaching the data requirements of these models have been well researched in the field. Several methods have been developed to aggregate datasets, or to adapt existing models trained on a larger, different dataset to a new smaller dataset. These methods are all aimed at developing the practical applicability of DL models in clinics and healthcare settings.

2.4 Learning from Multi-Source or Multi-Center Medical Datasets

Many researchers propose to combine datasets from different sources or centres to accommodate the size requirements of DL models as well as to improve generalizability by exposing the model to diverse data. Such approaches are primarily concerned with developing methods to accommodate the image-space differences [40, 65, 102], and do not consider annotation styles or their potential detriments to learning. In fact, there is an entire sub-field of deep learning called Federated Learning (FL) dedicated to learning strategies designed to leverage multiple datasets together without actually sharing any data. The field of FL primarily involves sharing learned parameters of a model to a pool instead of sharing data into a pool. Although this does address many practical concerns with data limitations and privacy issues, these methods often focus on standardization and harmonization and do not allow for the investigation of data factors that affect prediction outcomes or models [49, 81].

2.5 Domain Adaptation in Medical Imaging

Contrary to the previous section, Domain Adaptation (DA) approaches are aimed at directly addressing differences between datasets in DL approaches. Although they don't work towards aggregating or pooling data directly, these approaches do attempts to effectively leverage multiple datasets for efficient and practical training. DA works to adapt a given model trained on a *source* dataset to a new *target* dataset. The source dataset is the original dataset that the model was first trained on, and is often large and related to the target dataset somehow. The target dataset is the dataset required for the final desired task, and is often limited in size or constrained in some other way. The differences between the source and target datasets are often referred to as the domain shift. Domain shifts often result in severe performance drops of the algorithm before adaptation. Do-

main shifts can come in a variety of forms including class differences, synthetic-to-real adaptation, modality changes, or disease differences [6]. In this section, we review adaptation applications and methods that are most relevant to the goals of this thesis.

2.5.1 Image-Space Domain Shifts

Here we breifly highlight the existing works that aim to compensate for image acquisition shifts i.e., differences in the image space. Although not related to the focus of the paper, we quickly discuss these methods as image-based domain shifts are the primary concern of researchers in the field of medical imaging. These efforts focus on solutions that account for differences in images resulting from inconsistencies in system vendors, different types of image sequences, modalities, and acquisition protocols across sources [7, 38, 90, 92, 101]. Although these solutions are important for many medical datasets, these approaches do not address the changes in the annotation style.

2.5.2 Label-Space Domain Shifts

Many different factors can contribute to label-space domain shifts, as described in Sections 2.3.1 and 2.3.2. Although this field is not widely covered, there does exist a well-researched body of literature focused at the inter-rater variability or bias issue, as well as rater error issues. Although some papers in natural imaging semantic segmentation tasks address different class definitions resulting in label space shift [41, 51], this type of difference in label definition is not commonly investigated in medical segmentation datasets.

Inter-Rater Variability

As mentioned, several studies have focused on addressing the impact of inter-rater bias and the resultant annotation style. These studies primarily use datasets where multiple raters are each given the same patient cohort to segment. In these cases, differences in labels can be directly associated with the raters' opinion or style, experience level, or the

uncertainty of the target pathology [26, 35, 36, 94, 99]. This problem represents a specific scenario where the segmentation biases are found by collecting multiple annotations per sample. Schwartzman *et al* [80] showed that rater bias in training samples is actually amplified by neural networks. [50] showed that inter-rater biases can create problems for automated methods. They also noted that stochastic rater errors are more easily solved, and that consistent rater biases are the primary issue. This finding was supported by [47], where they concluded that for collaborative labelling efforts or datasets to be successful, you need unbiased labelers. Other researchers have also found that multiple labels with varying biases can provide important insights into uncertainty estimation [11, 94]. Inter-rater bias studies are an excellent example of how different annotations can effect medical DL networks; however, in many datasets, differences are not just attributed to one rater, and multiple annotations per image are often not available or feasible to collect.

Label Errors

Unlike in this thesis, many other works make the assumption that there does exist an attainable, absolute true ground-truth in medical segmentation. As previously discussed, this assumption is likely true for many natural imaging tasks, but due to the ambiguity in medical segmentation, this assumption may not always be accurate. In the works that make such assumptions in medical segmentation tasks, differences in annotations are often identified as label *errors*, or deviations from the truth. Other researchers refer to the aforementioned problem of inter-rater bias as a source of "label noise" [34,39,104], where differences between annotators can be described as noisy/erroneous variations around the real ground-truth. Some researchers have proposed noise-tolerant or noise-resistant losses or architectures to be robust to what error-corrupted labels in the target dataset of retinal segmentation tasks [105]. Vădineanu *et al* [91] performed an analysis on how different types of labelling errors in cell segmentation datasets can effect deep segmentation networks. They synthetically generate 3 different types of errors based on omission, bias, or inclusion and find that the U-Net is robust to most errors especially when error

severity is low. Though the findings of this paper are valuable in demonstrating the impact of isolated annotation variations on DL networks, the researchers do not address the complex annotation style differences investigated in this thesis.

2.5.3 Normalization-Based Adaptation Methods

The normalization layers in CNNs provide an important function, and recent research pioneered in the field of natural computer vision has shown the benefits of normalization layers in adaptation tasks. Conditional Instance Normalization (CIN) was first proposed in 2016 by Dumoulin *et al* [93], and was designed for artistic style transfer in image-space of natural images. In place of traditional instance normalization, CIN normalizes the activation, and applies a scale and shift learned from a style sample, as shown in Figure 2.8. By scaling and shifting the activations with style-based parameters, the authors were able to apply an artistic style, like that of a Monet painting, to any input image.

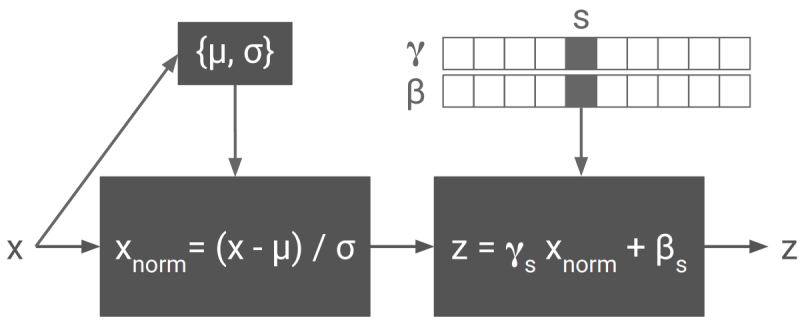


Figure 2.8: Conditional instance normalization mechanism, where x is the activation and γ_s and β_s are learned style parameters. Figure courtesy of [93].

Several other natural imaging papers have applied similar adaptive normalization-based methods to style transfer or image denoising problems [9,42,44,58,69]. Researchers have also applied such conditional normalization methods in medical problems focused on biases regarding the images [38] and relevant clinical information [31].

2.5.4 Fine-Tuning for Adaptation

Fine-tuning has been a long established training method used in DL in order to adapt an already trained network to new data. Fine-tuning a pre-trained network from a related or relevant dataset is especially beneficial in cases where not enough data is available to train a network from scratch. Fine-tuning rather than training from scratch can result in reasonable performance from very limited data. Some approaches fine-tune whole networks; however, in cases with limited data, fine-tuning fewer layers can help reduce overfitting [20, 20]. The question then follows: which layers to fine-tune? Generally, the consensus in the literature is that layers earlier in the network model low-level general changes like differences in image appearance. Layers deeper in the network model more detail-oriented features. [3] demonstrate this with ultrasound images. Since ultrasound images have large visual differences between datasets or tasks, fine-tuning earlier layers and then gradually fine-tuning later layers is a better approach. They were able to fine-tune with success from a network trained on many natural images, however they did require over 100 labelled samples for fine-tuning on the new task. [89] adapted a glioma segmentation network to data from new medical centers subject to both image-space and label space shifts. They tested a wide range of layer combinations for fine-tuning and found that fine-tuning one central layer in the encoder, and another central layer in the decoder was the optimal strategy for their conditions. Although the exact layers for fine-tuning may differ per-task, their general findings recommend that fine-tuning fewer layers is better with fewer samples, and that fine-tuning some layers in each portion of the network (encoder or decoder) is better than fine-tuning the entirety of only one half of the network. [88] instead propose an iterative layer-wise fine-tuning strategy for adapting a natural image network to medical tasks. The propose that the last layers are most significant in transfer learning tasks due to the specific features learned for each domain. They propose starting fine-tuning from the last layer, and incrementally including more layers in the fine-tuning process. The more different the target dataset is from the dataset used in the pre-trained network, the more early layers need be fine-tuned. [20] adapt a

U-Net trained for white matter hyperintensity segmentation to data acquired with different scanner resolutions, but consistent labelling processes. Due to the acquisition shift and the use of batch norm, which relies on moving average training statistics at test time, fine-tuning was necessary in order to adapt to image shifts. Since the tasks were the same, but the image specifics were different, they found that fine-tuning the later layers of the network lead to the best outcome.

In the context of annotation style shifts, fine-tuning is particularly useful as it exposes the network to the desired target annotation style. Especially in segmentation tasks, annotation styles can often not be quantified or described due to their nuances; therefore, by providing the network with examples of the annotation style, we give the network an opportunity to learn and generate the desired style. In the previously described works, annotation style shift is not an isolated problem, and other factors are also at play; however, these works still demonstrate the utility of fine-tuning. Given the previously demonstrated capability of CIN to model style information, we hypothesize fine-tuning only the CIN layers would be effective in adaptive a network to new annotation styles. Previously, [38] showed that fine-tuning only the source-specific batch norm parameters was successful in adapting a brain anatomy segmentation network to images from different medical centers. They experimented with fine-tuning from different source domains, and concluded that the batch norm parameters that resulted in the best performance (before fine-tuning) were the best parameters to fine-tune from. [42] proposed fine-tuning adaptive instance normalization layers to new domains in their denoising algorithm. They fine-tuned domain-specific instance normalization layers from old related domains to new target domains.

2.6 Summary

In this chapter, we provided the necessary background information on DL and medical image segmentation to understand the work of the thesis. Furthermore, we reviewed

literature that identified the sources of variability in the label space, and presented the limited research on domain shifts with label spaces. In doing so, we highlighted the shortcomings of current work and motivated the research in this thesis. There is a wide range of research on the effect and management of image-space shifts and label-space shifts caused by inter-rater bias. Previous research has shown just how important ground-truth labels are in DL networks, and has also demonstrated how many different factors can potentially affect those labels. Despite these influential studies, there is very limited research on how more complex and compounded variations in ground-truth annotations effect DL networks and their generalizability.

In clinical settings, inadequate DL models have potential to cause real medical consequences. Each clinical environment may have their own needs for their segmentation algorithm depending on the aim of the clinic, as described in the scenario posed in Figure 1.1a. However, due to the nature of DL models, they are unable to produce annotations in a style they have not seen before, therefore limiting their generalizability in a sense. As a result, they may produce segmentation predictions in a style inappropriate for a given clinic's goals. Recent work described in the previous sections has provided us with the tools to model such annotation styles, particularly, normalization-based adaptation methods. Although aspects related to annotation style shifts been addressed individually, including projects addressing inter-rater variability and population distribution shifts, researchers have not yet acknowledged the problem of annotation styles directly. As a result, this problem has gone largely unsolved. The next chapter will outline how to use CIN in a novel application to understand the impact of annotation styles and model them with DL networks.

Chapter 3

Conditional Instance Normalization for Modelling Annotation Styles

This chapter describes different methods used in this thesis to investigate and accommodate for differences in annotation styles and their various effects on DL segmentation. Specifically, an outline of the proposed method to model and adapt to different annotation styles in aggregated datasets is done. A general approach to investigate annotation styles using a series of different models is then described. This section also details our analysis method for discovering relationships between annotation styles using learned model parameters. Lastly, an explanation of the various approaches proposed for fine-tuning an existing network to a new annotation styles are defined.

3.1 Conditional Instance Normalization Method

Recall from Section 2.5.3, CIN is a style-adapting normalization mechanism. CIN layers are conditioned on auxiliary information to learn scale and shift parameters that can transform activations of the layer to obtain the desired style on the sample passed through. In this thesis, we propose a novel modification to an nn-UNet architecture by replacing traditional normalization layers with CIN layers to adapt the network to dif-

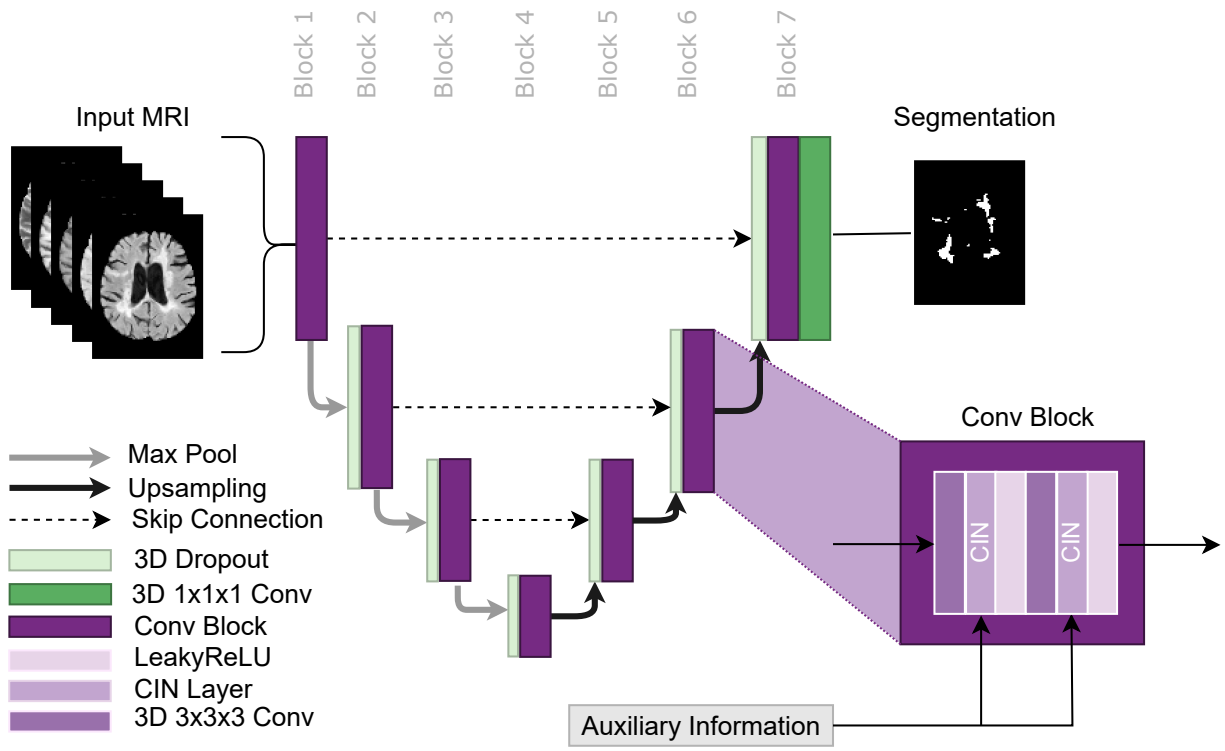


Figure 3.1: Left: Overview of modified nnUNet ([30]) architecture used to segment MS T2 lesions. Right: Detail of a conv block. It consists of a series of 3D 3x3x3 Convolution Layer, CIN layer, and a LeakyReLU activation layer.

ferent annotation styles for MS lesion segmentation as shown in Figure 3.1. Our method involves incorporating the CIN layer proposed by Dumoulin *et al* [93] into the target network to learn biases using a set of scale and shift parameters unique to each dataset or annotation style. This layer allows for dataset biases to be modelled while simultaneously performing the target task, as explained in Section 2.5.3. This thesis focuses on modelling dataset biases that lead to unique annotation styles. The CIN layer can be used in any network architecture where standard normalization layers are traditionally applied. Note that a conditional IN layer was used, rather than a conditional BN layer because BN relies on the mean and variance of the training set when performing normalization at test-time. This could result in undesirable behavior if the training statistics are not reflective of the testing statistics, and as such, BN approaches are not experimented with. The nn-UNet architecture is also modified to include dropout to help counter overfitting.

In our framework, the CIN layer works by scaling and shifting the normalized activations at every layer using a dataset-specific set of affine parameters during the forward pass. The CIN layer is represented by the following equation:

$$\mathbf{CIN}(\mathbf{z}) = \boldsymbol{\gamma}_s \left(\frac{\mathbf{z} - \boldsymbol{\mu}(\mathbf{z})}{\boldsymbol{\sigma}(\mathbf{z})} \right) + \boldsymbol{\beta}_s \tag{3.1}$$

where $\boldsymbol{\gamma}_s$ and $\boldsymbol{\beta}_s$ are the conditional affine parameters and where $\mu(\mathbf{z})$ and $\sigma(\mathbf{z})$ represent the per-channel mean and standard deviation of the input z, respectively. The affine parameters correspond to a specific dataset, s. The dataset-specific affine parameters are learned only from samples from the corresponding dataset. Other than the dataset-specific affine parameters in the CIN layers, all other network parameters are learned from all samples regardless of which dataset they came from. This allows the approach to leverage multiple datasets from different sources or annotation processes to learn common features while still taking into account dataset-specific annotation styles. A full system overview can be found in Figure 1.2.

3.2 Investigating the Impact of Annotation Styles

If different annotation styles exist in aggregated datasets, this thesis postulates that neglecting to account for them will result in a number of potentially detrimental outcomes. There are two main basic methods that do not consider annotation styles: 1) datasets are pooled together without any mechanism for style accounting, or 2) an individual model is made for each dataset. This thesis evaluates and highlights the downfalls of using both methods in the presence of annotation style shifts, which are referred to as Naive-Pooled and Single-Dataset models respectively. Naive-Pooled and Single-Dataset models are simple approaches that are used as baselines in this thesis, and they are not apart of the proposed methods of this thesis. These models are used only to compare and contrast with our proposed CIN-based method in the contexts of both practical implementations

and generalizability. Further details regarding the purposes of these baselines are described in the following Chapter 4.

3.2.1 Single-Dataset Models

Single-Dataset models are only trained on one dataset generated with only one singular annotation style. By having an independent model trained on only one dataset with one annotation style, the resultant model does allow for learning a specific annotation style. However, this approach requires a complete new model for every new dataset with a new annotation style. Each individual model would theoretically be able to model the annotation style of the dataset with the most accuracy, as these models have never seen any other annotation styles. Although this method does in a way accommodate and account for annotation styles by keeping separate models for each, this method is impractical. Storing all trained models separately takes up large amounts of storage space and is also difficult to implement in real work situations. As data continuously becomes available, there becomes a need to develop lifelong learning strategies that can be regularly built upon. And most importantly, these Single-Dataset models do not allow for learning from large, diverse, aggregated datasets. This is a particularly challenging limitation for medical computer vision since many medical datasets are very small, sometimes having even less than 50 patients for more rare diseases. Having the capacity to learn from multiple datasets collected under different protocols would be of great benefit for the medical field, allowing for larger training sets and more robust and reliable DL models. For this model, we use an identical model to that shown in Figure 3.1, but with a traditional IN layer instead of a CIN layer.

3.2.2 Naive-Pooled Models

For most problems in the DL field, big and diverse datasets with samples collected from many sources or situations will lead to a more generalizable model [23]. In this thesis,

this approach is referred to as the Naive-Pooled model, where datasets are pooled together and given to the model without any context of where the data came from. From the perspective of DL and generalizability, these types of models should be be more generalizable given that they had "seen" many more patient images obtained in different ways [96]. This model is used as a baseline to analyse the potential effects that annotation styles have on a DL network when they are gone unaccounted for. Specifically, a Naive-Pooled model is used to determine if different annotation styles across the datasets in the pool have consequences on the models ability to perform. This model will be given no auxiliary information about the data provided, and will effectively use only the image in an attempt to learn a singular annotation style that will satisfy all the different datasets used in training.

In certain situations, annotation styles may be predictable from the image. For example, if changes in scanner directly correlate with changes in annotation style, a Naive-Pooled method may be able to detect such pattern since the image is predictive of the annotation style. However, this is often not the case as some hospitals or research centers can share a scanner, or use the same scanner and still have different labelling processes. Furthermore, DL models will often require image normalization performed the preprocessing step to avoid image-based biases [82], and thus these images may no longer contain imaging information relevant to the labelling style.

Similar to the Single-Trial models, for this model, an identical model to that shown in Figure 3.1, but with a traditional IN layer.

3.3 Identifying Similar Annotation Styles Across Datasets

With the CIN method, each annotation style has its own set of IN parameters, once the model is fully trained, we have set of data that quantifies a style. With the CIN pipeline, we can explore how the learned dataset-specific scale and shift parameters of the fully-trained network can be used to determine unknown relationships between annotation

styles from the different sources. Understanding sources of annotation styles can not only improve future data collection methods and make for more fair performance evaluations, but it can allow for strategic pooling of compatible styles. Consider a scenario where each individual dataset on a specific disease is quite small (which is common for rare diseases), and each dataset is annotated independently with different annotation styles. In this scenario, if some datasets have fairly consistent annotation styles and data is limited, pooling them strategically and treating them as a *group* during training could be beneficial. By doing so, the style parameters in the CIN layer would have more examples to learn from, therefore creating more robust and well-informed affine parameter sets for each style *group*. Furthermore, by learning the potential factors that contributed to these different annotation styles, research centres can also consider changing their annotation methods in the future to make for easier data aggregation in later projects. An example of such factor would be the semi-automated labelling assistance software. If this software is found to be a considerable factor contributing to differences in styles across centres, the collaborating centres may consider using the same software in order to make aggregating their data easier in future studies.

To identify similarities in annotation style across datasets, I calculate the Cosine Similarity (normalized dot products) of the parameters between all sources for all CIN layers in the network in order to determine the relevant relationships and where they occur within the network. Specifically, the Cosine Similarity between two sources, s and $\tilde{s}$, is computed as follows:

$$\text{Cosine Similarity}_{scale_{s,\tilde{s}}}(n) = \frac{(\boldsymbol{\gamma}_{n_s} - \mathbf{1}) \cdot (\boldsymbol{\gamma}_{n_{\tilde{s}}} - \mathbf{1})}{|\boldsymbol{\gamma}_{n_s} - \mathbf{1}||\boldsymbol{\gamma}_{n_{\tilde{s}}} - \mathbf{1}|}$$

$$\text{Cosine Similarity}_{shift_{s,\tilde{s}}}(n) = \frac{\boldsymbol{\beta}_{n_s} \cdot \boldsymbol{\beta}_{n_{\tilde{s}}}}{|\boldsymbol{\beta}_{n_s}||\boldsymbol{\beta}_{n_{\tilde{s}}}|}$$

where n is the specific CIN layer in the network, and s and $\tilde{s}$ represent different sources. Since the scale parameters are initialized at 1, and a scale of 1 is representative of the

parameter having no effect on the activation, 1 is subtracted from all scale vectors before performing the Cosine Similarity calculation. This effectively relocates the origin to the initialization point of all scale parameters allowing for a more fair evaluation of the parameter changes that occurred during training.

By analysing Cosine Similarity between source parameters, high dimensional direction-based relationships between scale and shift parameter vectors of different sources can be quantified. To perform another analysis that now considers magnitude, a qualitative analysis of the linear norms of the CIN parameters is done. We attempt to look at the norms to detect any noticeable clusters or trends that may go undetected in the cosine similarity analysis.

3.4 Fine-Tuning to Images Labelled with New Annotation Styles

New data is constantly needed to improve the knowledge basis of a DL model and to learn potentially new phenomena. With new incoming data, datasets with new annotation styles are practically inevitable. Disease pathology definitions can evolve overtime, technology can advance, or new annotation goals my arise. Many factors can contribute to datasets coming in with new, unseen annotation styles thus enforcing a need for ways to continuously learn from data with new annotation styles.

One such way to continuously adapt models to data with new annotation styles is fine-tuning. The stance in this thesis, and supported by research [62, 93], is that the annotation style is able to be captured by the IN parameters of a network. As such, if a new dataset comes in and is image-normalized according to the established protocol from the pretrained model, it is possible that only IN parameters need be tuned in order to quickly learn new annotation styles. Regularizing the image space with the established protocol will minimize, if not eliminate, the image effects that contribute to the domain shift, leaving only the the annotation style shift. By only modifying the IN parameters and freezing

the rest of the network, one can learn from very few labelled samples without a major risk of overfitting. The benefit of this type of fine-tuning allows for efficient lifelong learning that reduces continual need for manual or semi-manual annotations. Researchers can provide a model with only a few labelled samples to give it an opportunity to learn the desired annotation style for the new task at hand.

There are two main ways to fine-tune a given network while only modifying only the IN layers. One can either: 1) re-initialize the IN layers and train from a scale value of 1 and a shift value of 0, or 2) initialize the IN layers at a pre-trained value of scale and shift. This thesis explores the benefits and differences of these two main fine-tuning strategies and how they affect performance and learned annotation style.

3.4.1 Fine-Tuning from No Affine

When training the original network, the IN parameters are initialized with a scale of 1 and a shift of 0, which effectively applies no affine transform initially. This allows the network to learn to apply an affine transformation where needed, from a starting point of *no affine* transformation. However, classical fine-tuning is often done by initializing a network or layer at the previously learned values. For style-specific parameters like IN parameters, this may be sub-optimal if the annotation style learned previously is not similar to the new one. Due to these reasons, this thesis proposes one fine-tuning strategy where all other non-IN layer parameters are kept frozen from the pre-trained model, and the IN parameters are initialized from *from no affine* with a scale of 1 and a shift of 0. This strategy can be performed in both Naive-Pooled models or CIN models. We term this strategy Fine-Tuning from No Affine.

3.4.2 Fine Tuning from Learned Affine

Fine-tuning from previously learned IN affine values may be sub-optimal if styles differ, but it could also potentially be beneficial if styles are similar. If a new incoming style is

known to be similar to a previously learned style, it may be beneficial to fine-tune from the IN parameters previously learned from the related style rather than from no affine (as previously described). This would initialize the IN parameters at already learned affine values that should theoretically also be similar to the ones needed for the new target style. Fine-tuning from learned affine parameters can be done in both Naive-Pooled or CIN models as well. However, for CIN models, the decision to fine-tune from learned values becomes more complicated. Since there are separate sets of IN affine parameters corresponding to each dataset for CIN-trained models, there are a variety of options for fine-tuning from learned parameters. This method becomes risky for CIN models as one must have knowledge about the annotation processes of the old and new datasets in order to make a informed decision about which set IN affine parameters to fine-tune from. In the case of Naive-Pooled models, there is also concern with the decision to fine-tune from learned affine values. If nothing is known about the annotation process of the original datasets, it may be difficult to decide if fine-tuning from learned affine values will be better than from no affine (scale 1 and shift 0). Furthermore, if the styles aggregated in the Naive-Pooled model were diverse, the IN parameters may not have converged at an optimal point and initializing from there may make the model worse-off when trying to learn a new, distinct style. Fine-tuning from learned affine parameters leads to an interesting question of whether the new style learned will even be true to the style of labels, as there will be influence from previously leaned datasets, even if the styles were similar. This question is investigated in detail in the final chapter of this thesis. The general approach of fine-tuning from pre-learned scale and shift values is referred to as Fine-Tuned from Learned Affine.

3.5 Summary

This chapter described the various methods that will be using throughout the rest of the thesis to both understand and model annotation styles. This thesis is conducted with a

U-Net backbone, with conditioned models having a CIN layer in place of the classical IN layer. Non-conditioned models include Naive-Pooled and Single-Trial models which will serve as a kind of baseline to compare conditioned models against, as well as to provide information on the impacts of annotation styles. This thesis will use our large collection of different MS clinical datasets in a series of extensive experiments with the described models in the following chapters to investigate annotation styles in pathological segmentation tasks with DL.

Chapter 4

Implementation and Experimental Details

This chapter provides details for the implementation and experimentation of the proposed methods in Chapter 3. An in-depth description of the datasets used for all experiments throughout the thesis is provided. This is then followed by an outline of the general experimental approach taken to investigate the impact of annotation styles using the proposed model in this thesis alongside the Single-Trial and Naive-Pooled models described in the previous chapter. Lastly, implementation and algorithm training specifics including hyperparameter optimization are described in brief.

4.1 Experimental Datasets

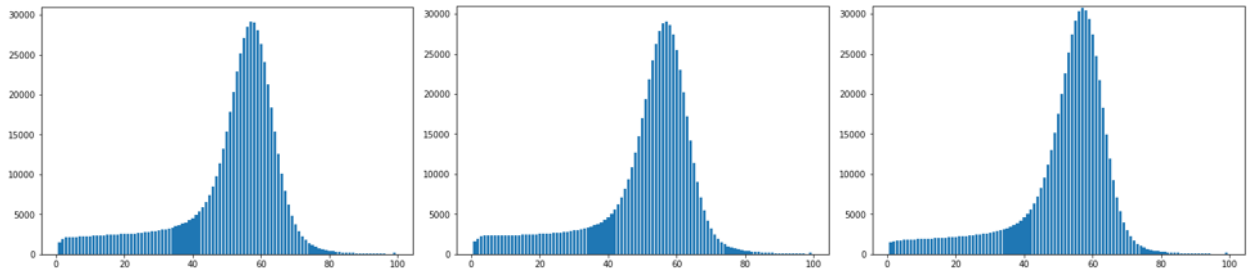
This thesis focuses on T2 lesion segmentation in MS clinical trial imaging datasets using the described nn-UNet. Each trial constitutes an individual datasets. The baseline scan of each patient for all trials is used, meaning that the images were taken before any treatment effects began. As a result, treatment effects are ruled out as a factor that effects annotation style in this thesis. The details of the trials used in this thesis are shown in Table 4.1.

Table 4.1: Trial names, disease phenotype, and the year of labelling for the trials and the corresponding trial code used to refer to all trials in the thesis.

Trial Name	Disease Phenotype	Trial Code	Year Labelled
BRAVO	Relapsing Remitting	RRMS-A	2008
DEFINE_ENDORSE		RRMS-B	2007
OPERA1		RRMS-C	2011
OPERA2		RRMS-D	2011
MAESTRO3	Secondary Progressive	SPMS-A	2007
ASCEND		SPMS-B	2011
ORATORIO	Primary Progressive	PPMS-A	2011
OLYMPUS		PPMS-B	2007

For all trials, each patient sample consists of 5 MR sequences acquired at 1mm $\times$ 1mm $\times$ 3mm resolution: T1-weighted, T1-weighted with gadolinium contrast, T2-weighted, Fluid Attenuated Inverse Recovery (FLAIR), and Proton Density. T2 lesion labels were generated at the end of each clinical trial, and were produced through an external process where trained expert annotators manually corrected a proprietary automated segmentation method. Since the trials were completed at different times, labels may have been generated with different versions of the automated segmentation method. Although different expert raters corrected the labels, there was overlap between raters across trials and all raters were trained to follow a similar labelling protocol. Each image only has one style of provided segmentation label (unlike in inter-rater bias studies). Furthermore, within one trial, the same labelling process was followed in order to keep the labels within the trial consistent. For these reasons, each of our clinical trials is believed has its own annotation style dataset for this thesis. Initially, data may have been preprocessed with different methods, but for these experiments, all data was re-processed from the native space (pre-spatial normalization or registration) using a consistent processing pipeline. This processing pipeline resulted the trials having nearly indistinguishable intensity distributions in the image space, as shown in the sample histograms in Figure 4.1. Although image-based biases were effectively normalized out with the re-processing, the image

effects would likely have influenced the labels during the labelling process through the semi-automated algorithm.



(a) SPMS-A histogram. **(b)** SPMS-B histogram. **(c)** RRMS-A histogram.

Figure 4.1: Sample intensity histograms of 3 different trials to demonstrate the image-space consistency between trials.

Throughout this thesis, different experiments use different combinations of trials from the described trial set in order to learn more about annotation styles and aggregating datasets. The datasets are balanced such that each trial consists of an equal number of patients, with 390 patients per trial. For all experiments, a 60/20/20 split is used for training, validation, and testing, respectively on a per-trial basis such that each experiment always has an equal number of trial samples in every stage of development. The same data split is used throughout all experiments in this thesis, meaning that the same patients are always in either training, validation, or testing regardless of the model or chapter of thesis. This allows for the most fair comparison between the different models trained throughout the entire thesis.

4.2 General Experimental Approach

As described in Section 3.2, Single-Dataset models (or Single-Trial models, as in this thesis) are used to first establish a type of baseline performance for any given trial. By only learning one annotation style, without influence from any other datasets, the Single-Trial model performance is considered in this thesis as an approximate for the *ideal* performance of any other network. As such, in this thesis, a Single-Trial model is trained for

every single trial described in Table 4.1. Next, we train a Naive-Pooled model for every aggregated dataset created in this thesis. This Naive-Pooled model will represent the performance resulting from the simplifying assumption that annotation styles will not effect the model performance. A Naive-Pooled model will be trained on every aggregated dataset that any CIN-based model will be trained on to allow for performance comparison across all experiments. Both the Naive-Pooled model and the Single-Trial model performance results are used to compare and contrast the performance of the proposed CIN-based method in the following chapters.

4.3 Performance Evaluation Metrics

In this thesis, T2 lesion segmentation performance will be evaluated with DICE and PR-AUC, as defined in Section 2.2. For DICE, the threshold used corresponds to the highest DICE score for each individual algorithm, and for each respective dataset: each algorithm will have 1 threshold for every dataset in the evaluation set, which corresponds to the maximum DICE performance on that dataset. For example, if an algorithm is trained and tested on Dataset 1, D_1, and Dataset 2, D_2, this specific algorithm will then have an associated $threshold_{D_1}$ and $threshold_{D_2}$. PR-AUC is also presented to provide a metric independent of threshold selection.

4.4 Implementation and Hyperparameter Optimization

Models in this thesis are developed with Python and PyTorch. All models in this thesis are trained using Binary Cross Entropy (BCE) Loss and Adam optimizer [43]. Random affine and random contrast augmentations are also used to prevent over fitting, and these transforms were implemented using the PyTorch API, MONAI. Each model is hyperparameter tuned independently depending on their validation performance. The learning rate, learning rate scheduler, dropout, and data augmentation hyper parameters were all

tuned on each individual model based on validation performance, specifically the validation loss. For the Single-Trial models, the hyperparameters that resulted in the best validation performance on the training trial were chosen, and this model was then tested against the other trial sets afterwards. For the multi-trial models, the hyperparameters that resulted in the best overall validation loss (across all trials were used). The range for the hyperparameter search is presented in Table 4.2

Table 4.2: Table detailing the hyperparameter search space for all models.

Hyperparameter	Sub-Parameter	Values
Epochs	-	150-250
Learning Rate	-	1e-4 to 3e-4, with 1e-5 weight decay
Scheduler	Epochs	[25, 50, 100, 150, 200] to [200]
	Gamma	1/3 to 1/2
Random Affine	Probability	0.5-0.8
	Rotate Range	4(pi/180)-8(pi/180) for all dimensions
	Shear Range	0.08-0.10 for all dimensions
	Scale Range	0.08-0.10 for all dimensions
Random Contrast	Probability	0.5-0.8
	Gamma	(3/4, 4/3) and (2/3, 3/2)

4.5 Summary

This chapter presented details for all the experiments in the following chapters, including the data to be used, the general experimental approach using Naive-Pooled, Single-Trial, and CIN-based models, and further implementation details. These details provide important information for any other researchers hoping to replicate the results of this thesis. The following chapters will apply these models and experimental approaches in order to investigate and account for annotation styles in the described MS T2 lesion segmentation datasets.

Chapter 5

Generalizability and the Impact of Annotation Styles

In Chapter 2, how annotation styles can arise, and how they can affect the training and evaluation of neural networks was described. In this chapter, the proposed CIN-based segmentation method from Section 3.1 will be evaluated. First, this chapter investigates how the different segmentation models handle a varied, multi-phenotype aggregated dataset. Since annotation protocols and software are both kept consistent for the entirety of a single trial (independent of disease phenotype), this chapter first investigates annotation style shifts between individual trials. In the first portion of this chapter, to account for annotation style differences between trials in the multi-phenotype dataset, the proposed CIN-based model is conditioned on the trial from which the patient came. This model is referred to as the *Trial-Conditioned model*. Naive-Pooled models and Single-Trial models are also studied in order to determine potential effects that annotation styles have on generalizability and overall model performance. The Trial-Conditioned model is evaluated for its ability to model and accommodate different annotation styles from different trials in one singular model, and the benefits of such functionality are assessed. The proposed Trial-Conditioned model also has the capacity to produce segmentation predictions in any learned annotation style, as explained in Section 3.1. By having segmentation results for

each learned annotation style on any given sample, we can study the differences in style both qualitatively and quantitatively in this chapter. An analysis of the different styles produced by the Trial-Conditioned model is also used to challenge the common assumptions around the generalizability of DL algorithms. This chapter explores the benefits of the proposed model for accounting for annotation styles in practical model implementations with a varied dataset.

Next, this chapter investigates the possible confounding factor that is the disease phenotype. As described in Chapter 1, MS has 3 main disease phenotypes. These disease subtypes can have various effects on lesion load or new lesion development. Despite the fact that the different trial datasets described in Section 4.1 are heavily normalized across the image space, and are labelled with relatively similar versions of annotation pipelines, the disease phenotype is not consistent across all trials. During the semi-automated annotation process, raters often have access to additional medical information, including disease phenotype. Since trained expert raters have knowledge and certain expectations for a given disease phenotype, this can result in the annotations being affected by *observer bias*. Observer bias is the systemic discrepancy from the truth during recording of data due to external influence. Although "truth" in the context of MS lesion segmentation isn't truly attainable, additional medical information such as phenotype could still lead to consistent differences in annotation style. For example, if the rater knows that a patient has a progressive type of MS, they may be expecting to see higher lesion volumes and thus look for them more carefully or become aware of them more easily. Furthermore, many researchers do divide their datasets based on diseases or pathology specifics. Therefore, in this chapter, we also aim to investigate some of these assumptions using CIN-based, Naive-Pooled, and Single-Trial models with a focus on disease phenotype. The chapter continues to conduct experiments with these models using *phenotype-consistent* aggregated datasets. Furthermore, in order to understand the relative impact, if any, that disease phenotype has on annotation style compared to individual trial identity, we also

investigate the impact of conditioning on phenotype compared to conditioning on the trial to enhance the understanding of the factors that can and may effect annotation styles.

5.1 Generalizability and Annotation Styles

In this section, we use our aforementioned models, including Single-Trial, Naive-Pooled, and Trial-Conditioned models to show the challenges that annotation styles can cause for generalizability and evaluation standards in diverse MS T2 lesion segmentation datasets.

5.1.1 Experiment Details

In the first part of this chapter, we use a Trial-Conditioned model trained on RRMS-A, RRMS-B, SPMS-A, SPMS-B, PPMS-A, and PPMS-B. We also compare to a Naive-Pooled model trained on the same data, as well as the corresponding Single-Trial baselines. Since the CIN models have a different set of parameters per trial, we are also able to see how a style from one trial compares to a style from another trial. We do so by passing in a test sample from one trial, say RRMS-A, and telling the model to label it as though it was SPMS-B. We then compare this predicted segmentation to the "ground truth" from RRMS-A and we perform an analysis on qualitative results as well as quantitative performance metrics. We do so for all trial styles and trial data combinations in order to explore annotation styles. The results from this series of experiments also serve to demonstrate the existence of annotation styles and subjective lesion definitions.

5.1.2 Generalizability Results and Discussion

This section presents the qualitative and quantitative results of the aforementioned models, as well as a discussion and analysis of the results.

Single-Trial Model Evaluation

This section shows the performance of the Single-Trial models for all trials used in the later Trial-Conditioned model. These baselines each represent a model that fully and only learns one isolated style, and as such, they offer valuable insights on annotation style and generalizability.

Table 5.1: Performance on test sets: experiments on Single-trial models.

(a) F1 Performance on test sets.

	Model	Training Set	Test Performance (F1)					
			RRMS-A	RRMS-B	SPMS-A	SPMS-B	PPMS-A	PPMS-B
1	**Single-Trial**	RRMS-A	**0.784**	0.779	0.776	0.681	0.647	0.646
2	**Single-Trial**	RRMS-B	0.775	0.779	0.766	0.678	0.638	0.646
3	**Single-Trial**	SPMS-A	0.778	**0.785**	**0.782**	0.679	0.645	0.647
4	**Single-Trial**	SPMS-B	0.691	0.689	0.686	**0.730**	**0.709**	0.693
5	**Single-Trial**	PPMS-A	0.697	0.699	0.690	**0.731**	0.699	0.681
6	**Single-Trial**	PPMS-B	0.669	0.671	0.663	0.682	0.646	**0.742**

(b) PR-AUC performance on test sets.

	Model	Training Set	Test Performance (PR-AUC)					
			RRMS-A	RRMS-B	SPMS-A	SPMS-B	PPMS-A	PPMS-B
1	**Single-Trial**	RRMS-A	**0.875**	0.865	0.866	0.750	0.707	0.704
2	**Single-Trial**	RRMS-B	0.864	0.866	0.854	0.746	0.692	0.706
3	**Single-Trial**	SPMS-A	0.869	**0.874**	**0.873**	0.751	0.704	0.705
4	**Single-Trial**	SPMS-B	0.773	0.764	0.764	**0.818**	**0.786**	0.766
5	**Single-Trial**	PPMS-A	0.783	0.782	0.773	**0.818**	0.773	0.749
6	**Single-Trial**	PPMS-B	0.746	0.744	0.734	0.746	0.700	**0.828**

The generalization of Single-Trial baselines to other trials demonstrates a part of the problem: when a model only learns to produce one annotation style, comparing the predictions of said model to the ground-truths of other trials (with other annotation styles) results in poor performance. These models show a clear performance degradation of up to 10% when applied to different trial test sets. For example, if we look at the RRMS-A performance on the RRMA-A test set, we see a very reasonable performance of 0.784 F1 and 0.875 PR-AUC; however, when applying this model to PPMS-B, we only see a 0.646 F1 and a 0.704 PR-AUC. Without acknowledging the issue of annotation style, a researcher might assume that the RRMS-A model has "poor" generalizability and is, as a result, a

bad model. In reality, its more nuanced than that, and likely the RRMS-A model is under performing on PPMS-B because the annotation processes or protocols varied between the two ground truths of the datasets. This explicitly shows the issue of comparing models trained on one annotation style to ground truths generated in a conflicting annotation style. Since we only have one annotation style per trial, it is essentially impossible to accurately and fairly evaluate the generalizability of a single-trial model on another trial with a different annotation style.

Trial Conditioning Results

In this section, experiments are conducted to examine the CIN model and its benefits in situations with variable annotation styles, specifically in the context of generalizability. Furthermore, we compare our CIN Trial-Conditioned model to a Naive-Pooling baseline to highlight more challenges caused by annotation styles.

Comparisons of the F1 and PR-AUC segmentation scores for the Single-Trial models, Naive-Pooled model, and the proposed Trial-Conditioned model are shown in Table 5.2. The Naive-Pooling model results demonstrate another consequence of ignoring variable annotation styles. Many studies use Naive-Pooling across several datasets in order to increase the dataset size in the hopes of improving performance and generalizability. However, the results show that when training on more data but with different annotation styles without context or consideration (as done in the Naive-Pooling model), the model is unable to perform well in any annotation style, effectively eliminating the expected benefit of using more data. This is likely because it is unable to successfully learn any annotation style, and when tested on the trials, it cannot produce the required label and thus appears to fail to generalize. Naive-pooled models have no information regarding the different requirements for segmentation across datasets in the pool and therefore have no way of predicting the right annotation style to produce. On the other hand, provided with context on the source of each patient sample, the Trial-Conditioned model shows performance on par with the Single-Trial models. This illustrates CIN's ability to learn

Table 5.2: Performance on test sets for the Trial-Conditioned model, Naive-Pooling model, and Single-Trial models. The Trial-Conditioned model is passed the trial ID of the sample during both training and test time.

(a) F1 performance on test sets.

	Model	Training Set	Testing Set (metric F1)					
			RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	**Single-Trial**	RRMS-A	0.784	-	-	-	-	-
2	**Single-Trial**	RRMS-B	-	0.779	-	-	-	-
3	**Single-Trial**	SPMS-A	-	-	0.782	-	-	-
4	**Single-Trial**	SPMS-B	-	-	-	0.730	-	-
5	**Single-Trial**	PPMS-A	-	-	-	-	0.699	-
6	**Single-Trial**	PPMS-B	-	-	-	-	-	0.742
7	**Naive-Pooling**	**All**	0.766	0.754	0.756	0.722	0.700	0.738
8	**Trial-Conditioned**	**All**	**0.787**	**0.789**	**0.788**	**0.737**	**0.709**	**0.744**

(b) PR-AUC performance on test sets.

	Model	Training Set	Testing Set (metric PR-AUC)					
			RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	**Single-Trial**	RRMS-A	0.875	-	-	-	-	-
2	**Single-Trial**	RRMS-B	-	0.866	-	-	-	-
3	**Single-Trial**	SPMS-A	-	-	0.873	-	-	-
4	**Single-Trial**	SPMS-B	-	-	-	0.818	-	-
5	**Single-Trial**	PPMS-A	-	-	-	-	0.773	-
6	**Single-Trial**	PPMS-B	-	-	-	-	-	0.828
7	**Naive-Pooling**	**All**	0.859	0.843	0.847	0.809	0.776	0.822
8	**Trial-Conditioned**	**All**	**0.879**	**0.878**	**0.879**	**0.826**	**0.786**	**0.829**

the trial-specific annotation styles all with only one model. This is especially useful when datasets are collected on the same or similar pathologies, but with different clinical goals in mind when generating the labels. The final Trial-Conditioned model is also capable of producing any annotation style on any output, thus providing researchers with multiple "opinions" on a given sample, as shown in Figure 5.1.

Figure 5.1 shows the results for a test sample from one trial (RRMS-B) as segmented by the Trial-Conditioned model using the CIN parameters from all the different trials. The results clearly demonstrate unique segmentation styles across trials. One observation that can be made is that conditioning on the SPMS-B, PPMS-A, and PPMS-B trials results in a noticeable relative over-segmentation on some of the larger lesions in all three cases. However, these styles cannot be reduced to simple over-segmentation. Using these three

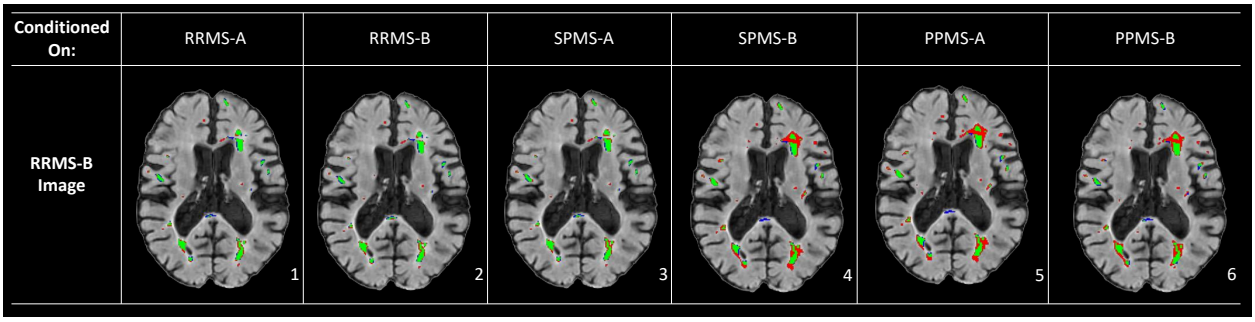


Figure 5.1: RRMS-B example labelled with the Trial-Conditioned model using different annotation styles. Green is true positive, red is false positive, and blue is false negative with respect to the RRMS-B "ground truth" label.

Table 5.3: Performance on test sets for the Trial-Conditioned model using the different annotation styles.

(a) F1 performance on test sets.

	Conditioning Style	Testing Set (metric F1)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	RRMS-A	**0.787**	**0.789**	**0.786**	0.684	0.650	0.667
2	RRMS-B	0.786	**0.789**	**0.787**	0.683	0.652	0.671
3	SPMS-A	0.784	**0.789**	**0.788**	0.682	0.650	0.669
4	SPMS-B	0.708	0.703	0.697	**0.737**	**0.712**	0.703
5	PPMS-A	0.718	0.716	0.707	**0.737**	0.709	0.693
6	PPMS-B	0.694	0.699	0.690	0.702	0.664	**0.744**

(b) PR-AUC performance on test sets.

	Conditioning Style	Testing Set (metric PR-AUC)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	RRMS-A	**0.879**	**0.878**	**0.878**	0.761	0.717	0.737
2	RRMS-B	0.878	**0.878**	**0.878**	0.760	0.718	0.742
3	SPMS-A	0.876	**0.878**	**0.879**	0.759	0.717	0.740
4	SPMS-B	0.796	0.783	0.781	**0.826**	**0.791**	0.781
5	PPMS-A	0.808	0.801	0.794	**0.826**	0.786	0.772
6	PPMS-B	0.781	0.783	0.775	0.782	0.735	**0.829**

label styles also results in completely missing the lesion in the posterior part of the white matter tract located in the brain's midline. The similarity in the results across the SPMS-B, PPMS-A, and PPMS-B segmentation maps suggests that they form a potential annotation

style subgroup. On the other hand, the RRMS-A and SPMS-A segmentation styles are similar to that of RRMS-B and therefore result in very accurate results according to the RRMS-B label. This suggests that these three styles make up another subgroup. These groupings do show some commonalities between phenotype, as both of the PPMS-A and PPMS-B trials seem to have similar annotation styles, and the same goes for RRMS-A and RRMS-B. These relationships will be further investigated in the next section, 5.2.

In addition to qualitative results, the relationship between these sets of trials (SPMS-B, PPMS-B, PPMS-A) and (RRMS-A, SPMS-A, RRMS-B) is also demonstrated in the quantitative results presented in Table 5.3. This table shows the performance on the test set for the conditioned model using the different trial styles (ie conditioning on the wrong trial variable at test time). These results show that the best SPMS-B performance is obtained not only by the SPMS-B style, but also by the PPMS-A style. Additionally, the best RRMS-B performance is achieved by the RRMS-B, SPMS-A, and RRMS-A styles. The best PPMS-A performance is achieved by the SPMS-B style. Aside from the commonalities, we can also see that certain styles are rather conflicting. The SPMS-A CIN parameters yield very good performance on RRMS-A, RRMS-B, and SPMS-A; however, they obtain subpar performance on all other trials by up to 5%. These results further stress the importance of considering annotation styles when evaluating the generalizability of a trained model. Models trained on a dataset with an annotation style that is compatible with that of test set may appear more generalizable than a model trained on a dataset with a conflicting annotation style. Thus annotation styles lead to confusing and potentially incorrect conclusions being drawn regarding model performance if they are not correctly accounted for.

5.2 Investigating the Impact of Phenotype on Annotation Styles

This section focuses on the potential impact that disease phenotype may have on annotation styles. Specifically, this section conducts a series of experiments to determine the potential impact that observer bias caused by knowledge of disease phenotypes has on annotation styles and DL networks. This section uses Naive-Pooled, Single-Trial, and Trial-Conditioned models as in the previous section, but focuses on phenotype-consistent aggregated datasets. Additionally, more experiments are conducted with the mixed-phenotype dataset from the previous chapter.This dataset is used with another CIN-based model conditioned on the disease phenotype, referred to as the *Phenotype-Conditioned* model.

5.2.1 Experiment Details

We create three different aggregated datasets from the previously described set in Section 4.1: RRMS-only trials, SPMS-only trials, and PPMS-only trials. Each dataset was divided into training, validation, and testing as per the protocol described in 4.1. For example, for the SPMS experiments, the final training set would consist of 234 SPMS-A patients and 234 SPMS-B patients (60% of 390 patients) totalling a training set of 468 patients.

By training a Naive-Pooled and a Trial-Conditioned model on phenotype-consistent datasets and comparing the performance to single-trial baselines, this series of experiments is aimed at determining if relationships exist between the annotation styles of datasets with the same phenotype. Lastly, we compare performance of a Trial-Conditioned model to a Phenotype-Conditioned model trained on the previously used mixed-phenotype dataset. This comparison will inform us of the relative influence of phenotype compared to individual trial collection processes on annotation style.

5.2.2 Results and Discussion

This section presents test performance results for the RRMS-only, SPMS-only, and PPMS-only dataset experiments, as described in the previous section. We also perform a discussion on the results.

Annotation Styles in RRMS-Only Datasets

Table 5.4: Performance on the test sets of the RRMS-only experiments.

(a) DICE performance on test sets.

	Model	Training Set	Test Performance (DICE)			
			RRMS-A	RRMS-B	RRMS-C	RRMS-D
1	**Single-Trial Baseline**	RRMS-A	**0.784**	**0.779**	0.664	0.688
2	**Single-Trial Baseline**	RRMS-B	0.775	**0.779**	0.657	0.682
3	**Single-Trial Baseline**	RRMS-C	0.701	0.694	**0.712**	**0.729**
4	**Single-Trial Baseline**	RRMS-D	0.708	0.702	0.700	**0.732**
5	**Naive-Pooled**	RRMS-A, B, C, D	0.737	0.735	**0.709**	**0.731**
6	**Trial-Conditioned**	RRMS-A, B, C, D	**0.779**	**0.777**	**0.718**	**0.737**

(b) PR-AUC performance on test sets.

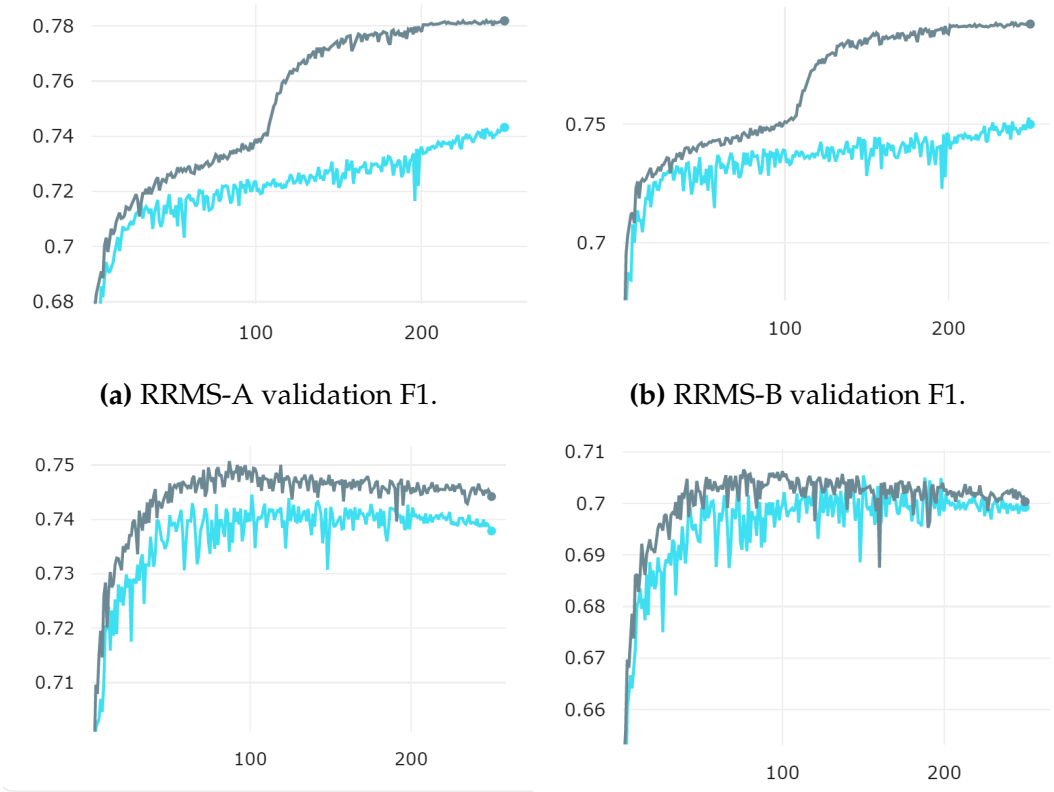
	Model	Training Set	Test Performance (PR-AUC)			
			RRMS-A	RRMS-B	RRMS-C	RRMS-D
1	**Single-Trial Baseline**	RRMS-A	**0.875**	**0.865**	0.728	0.758
2	**Single-Trial Baseline**	RRMS-B	0.864	**0.866**	0.723	0.750
3	**Single-Trial Baseline**	RRMS-C	0.788	0.776	**0.793**	**0.809**
4	**Single-Trial Baseline**	RRMS-D	0.797	0.782	0.774	**0.813**
5	**Naive-Pooled**	RRMS-A, B, C, D	0.829	0.824	0.791	**0.810**
6	**Trial-Conditioned**	RRMS-A, B, C, D	**0.870**	**0.865**	**0.802**	**0.818**

Table 5.4 shows the performance of all the RRMS Single-Trial baselines on all RRMS trials in rows 1-4. These results clearly show that there is still an annotation style shift between these 4 datasets, despite the fact that all images were acquired from patients with the same disease phenotype. This is demonstrated by the major performance gaps between the trials for any given Single-Trial baseline. For example, the RRMS-A Single-Trial model does very well not only on the RRMS-A dataset (as expected), but also on the RRMS-B test set, effectively demonstrating its capacity to perform well on data it has

not seen. However, performance on RRMS-C and RRMS-D is about 5% below the respective RRMS-C and RRMS-D Single-trial model performance. This relative performance degradation is likely caused by annotation style differences between the two datasets, given that their intensity distributions are approximately equal, the pathology is the same, and the phenotype is also the same. The only major differences remaining between the datasets is the annotation protocol, and corresponding semi-automated software.

Te problem of annotation shift is also demonstrated by the poor performance of the Naive-Pooled model, and the benefit of Trial-Conditioned models when trained on all 4 RRMS datasets. When pooling together datasets with different annotation styles, we see quite a performance degradation on some trials relative to their respective Single-Trial baseline performance. Furthermore, by conditioning the model on the trial identity, performance on all trials improves, and is at least on-par with Single-Trial baselines. This indicates that the trial identity is an important factor in generating appropriate segmentation predictions for MS T2 lesions, despite the consistent phenotype. RRMS-A and RRMS-B in particular suffer significant performance degradation from the Naive-Pooled model relative to their Single-Trial baselines, but this degradation is not present in the results from the Trial-Conditioned model. This trend can, in part, be explained by Figure 5.2. The RRMS-A and RRMS-B training curves both show important differences between the Naive-Pooled model and the Trial-Conditioned model: the Trial-Conditioned model undergoes *double descent* [57], and the Naive-Pooled model does not. Double descent refers to the phenomenon of performance improving, plateauing, and then undergoing another wave of improvement [57]. After epoch 100, both the RRMS-A and RRMS-B validation DICE score for the Trial-Conditioned model experiences another large jump in performance. This trend suggests that RRMS-A & B may have a more complex annotation style, meaning that there could be more small lesions or more irregular lesion borders labelled in these datasets. This could be caused by the specific goals or protocols of the different clinical trials. By providing the trial identity, the Trial-Conditioned model is likely able to achieve double descent and accommodate those more complex styles. RRMS-C & D

on the other hand may have relatively simpler styles, so when the model is provided with no context of the dataset source, as in the Naive-Pooled model, it may not be able to converge to the more complex style of A & B, and instead learns a more simple style similar to C & D. This hypothesis may also be supported by the reasonable performance of the Naive-Pooled model on C & D. The similarity between RRMS-A and RRMS-B was previously identified in Section 5.1.2; however, these experiments have clarified that this commonality is likely not caused by the common disease phenotype, given that RRMS-C and D appear to be labelled with different annotation styles. The similarity between RRMS-A and RRMS-B are more likely due to the aforementioned factor of the annotation protocols used for these trials.



(a) RRMS-A validation F1.

(b) RRMS-B validation F1.

(c) RRMS-C validation F1.

(d) RRMS-D validation F1.

Figure 5.2: Validation F1 vs training epoch curves during training for RRMS-only experiments for Trial-Conditioned (grey) and Naive-Pooled (blue) models.

Annotation Styles in SPMS-Only Datasets

Table 5.5: Performance on the test sets of the SPMS-only experiments.

(a) DICE performance on test sets.

	Model	Training Set	Test Performance (DICE)	
			SPMS-A	SPMS-B
1	**Single-Trial Baseline**	SPMS-A	**0.782**	0.679
2	**Single-Trial Baseline**	SPMS-B	0.686	**0.730**
3	**Naive-Pooled**	SPMS-A & B	0.723	**0.727**
4	**Trial-Conditioned**	SPMS-A & B	0.696	**0.734**

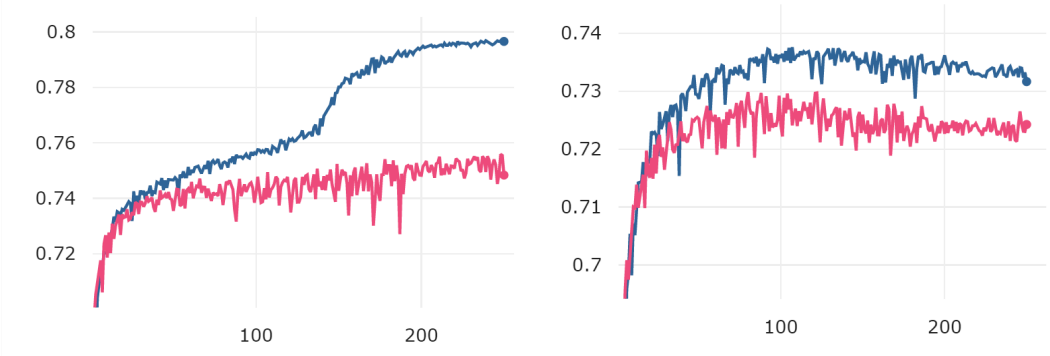
(b) PR-AUC performance on test set.

	Model	Training Set	Test Performance (PR-AUC)	
			SPMS-A	SPMS-B
1	**Single-Trial Baseline**	SPMS-A	**0.873**	0.751
2	**Single-Trial Baseline**	SPMS-B	0.764	**0.818**
3	**Naive-Pooled**	SPMS-A & B	0.813	**0.814**
4	**Trial-Conditioned**	SPMS-A & B	0.778	**0.821**

Similarly to Section 5.2.2, this section first presents the existence of an annotation style shift in the SPMS-consistent datasets in Table 5.5. The Single-Trial SPMS-A model severely under performs on the SPMS-B test set and vice versa, indicating a probable change in annotation style. With respect to the comparison between the models trained on the aggregated SPMS dataset, we also see similar trends to the RRMS results. However, the difference between Naive-Pooled and Trial-Conditioned models is not so clear.

Previously, we saw that the Trial-Conditioned model was fairly consistent with the Single-Trial model performance, and generally an improvement on the Naive-Pooled model. In the SPMS experiments, we see that the Trial-Conditioned model noticeably under performs the Naive-Pooled model for SPMS-A. SPMS-A is also a trial that experiences double descent for the Trial-Conditioned model (shown in Figure 5.3), as in RRMS-A and RRMS-B, suggesting that it too may be a more complex annotation style. The interesting finding in this series of experiments is that even though the Trial-Conditioned model outperforms the Naive-Pooled model by 5% on the validation set for SPMS-A, this performance benefit is completely diminished on the test set. There is still a slight per-

formance improvement from the Trial-Conditioned model relative to the Naive-Pooled model on SPMS-B, as consistent with the validation results. This performance degradation on SPMS-A for the Trial-Conditioned model could potentially be due to the fact that there are only 2 different trials in this dataset with two distinct and possibly very conflicting styles. As a result, the Trial-Conditioned model may be unable to learn both shared parameters as well as independent CIN parameters while converging on the complex SPMS-A style. This incompatibility of SPMS-A and SPMS-B corroborates the previous grouping identified in Section 5.1.2, and further supports the possibility that observer bias caused by knowledge of disease phenotype may be overshadowed by other labelling protocol differences.



(a) SPMS-A validation F1.

(b) SPMS-B validation F1.

Figure 5.3: Validation F1 vs training epoch curves during training for SPMS-only experiments for Trial-Conditioned (navy) and Naive-Pooled (pink) models.

Annotation Styles in PPMS-Only Datasets

The PPMS Single-Trial baselines demonstrate a similar annotation style issue with each Single-Trial baseline struggling to annotate the other trial, as shown in Table 5.6. The experiments on the aggregated PPMS dataset also further prove the annotation style shift, as well as demonstrate the benefit of using Trial-Conditioned models. The benefit of Trial-Conditioned models over Naive-Pooled models may be more clear in this experiment series than in the SPMS series due to relatively more similar annotation styles between

Table 5.6: Performance on the test sets of the PPMS-only experiments.

(a) F1 performance on test sets.

	Model	Training Set	Test Performance (DICE)	
			PPMS-A	PPMS-B
1	**Single-Trial Baseline**	PPMS-A	**0.699**	0.681
2	**Single-Trial Baseline**	PPMS-B	0.646	**0.742**
3	**Naive-Pooled**	PPMS-A&B	**0.702**	0.739
4	**Trial-Conditioned**	PPMS-A&B	**0.699**	**0.744**

(b) PR-AUC performance on test sets.

	Model	Training Set	Test Performance (PR-AUC)	
			PPMS-A	PPMS-B
1	**Single-Trial Baseline**	PPMS-A	**0.773**	0.749
2	**Single-Trial Baseline**	PPMS-B	0.700	**0.828**
3	**Naive-Pooled**	PPMS-A&B	**0.778**	**0.825**
4	**Trial-Conditioned**	PPMS-A&B	**0.770**	**0.830**

the PPMS-A and PPMS-B styles. Although they are different, as exemplified by the previously discussed results, neither PPMS-A nor PPMS-B undergo double descent, shown by Figure 5.4 (as seen in RRMS-A, RRMS-B, and SPMS-A). This could indicate that they are slightly more similar styles and therefore the Trial-Conditioned model can converge to reasonable trial-specific CIN parameters as well as shared parameters with only 2 trials. This is also supported by the previous groupings in Section 5.1.2 that identified PPMS-A and PPMS-B as having similar annotation styles.

Comparing Impact of Phenotype to Trial on Annotation Styles

Lastly, a comparison is done between the relative importance of phenotype and trial identity on annotation style in mixed-phenotype datasets. In practice, this will allow for more data to train a single model. Although MS phenotypes may have different disease courses, images from patients of different disease phenotype still provide valuable information for lesion segmentation. Comparison between the Phenotype -Conditioned, Trial-Conditioned, and Naive-Pooled models is shown in Table 5.7. Although both Trial-Conditioned and Phenotype-Conditioned models offer some improvement to the Naive-

(a) PPMS-A validation F1.

(b) PPMS-B validation F1.

Figure 5.4: Validation F1 vs training epoch curves during training for PPMS-only experiments for Trial-Conditioned and Naive-Pooled models.

Table 5.7: Comparison between Trial- and Phenotype-Conditioning and Naive pooling in mixed phenotype datasets.

(a) F1 performance on test sets.

	Model	Test Performance (DICE)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	**Naive-Pooled**	0.766	0.754	0.756	0.722	**0.700**	**0.738**
2	**Phenotype-Conditioned**	**0.778**	0.777	0.744	0.725	**0.707**	**0.740**
3	**Trial-Conditioned**	**0.787**	**0.789**	**0.788**	**0.737**	**0.709**	**0.744**

(b) PR-AUC performance on test sets.

	Model	Test Performance (PR-AUC)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	**Naive-Pooled**	0.859	0.843	0.847	0.809	0.776	**0.822**
2	**Phenotype-Conditioned**	**0.870**	0.867	0.832	0.813	**0.785**	**0.826**
3	**Trial-Conditioned**	**0.879**	**0.878**	**0.879**	**0.826**	**0.786**	**0.829**

Pooled model, the Trial-Conditioned model stands out. Particularly for SPMS-A, there is a significant improvement from Phenotype-Conditioned to Trial-Conditioned. This supports the earlier hypothesis that SPMS-A and B have particularly different annotation styles, as forcing them to share one set of CIN parameters (as in the Phenotype-Conditioned model) results in a decrease in performance relative to both Trial-Conditioned *and* Naive-Pooled models. These findings indicate that, although some trials with patients

of the same disease phenotype may have similar annotation styles, one cannot assume that this is true for all trials and all disease phenotypes.

5.3 Summary

In this section, traditional notions of generalization were challenged in the context of supervised medical image pathology segmentation models. This section poses that poor generalization performance of automated methods across datasets may partially be a consequence of differing annotation styles, rather than solely the result of scanner/site differences as is often assumed. The results from this section indicate that differences in annotation styles can arise due to a number of factors, specifically embedded in the annotation process or protocol, and cannot be simplified to differences in acquisition parameters, preprocessing, or disease phenotype.

Given the many factors that can effect the annotations of a given dataset, this chapter demonstrates that annotation biases are essentially unavoidable. This chapter shows that even datasets with image normalization preprocessing, no treatment effects, *and* consistent disease phenotypes still have annotation style shifts. As such, this thesis proposes to model them and accommodate them. This chapter clearly shows the benefit of the proposed CIN-based Trial-Conditioned model in the context of effectively leveraging multiple datasets by showing the comparable performance with both Single-Trial and Naive-Pooled models. The differences between the annotation styles have proven to be distinct but complex, and despite this, the scaling/shifting learned by CIN parameters are able to capture these unique styles in an easy-to-implement normalization layer. The unique functionality of the proposed conditioning framework not only allows for more efficient use of available data, but also provides both researchers and clinics with a model that is capable of producing multiple tailored outputs for specific sub-tasks.

Chapter 6

Identifying Datasets with Similar Annotation Styles for Strategic Aggregation

In the previous chapter, several results pointed towards similar annotation styles between trials, particularly (SPMS-B, PPMS-B, PPMS-A) and (RRMS-A, SPMS-A, RRMS-B). Given the slight trend of trials of the same disease phenotype tending to group together, further analysis was conducted to uncover the impact of disease phenotype on annotation style. The results from this investigation concluded that trials of the same disease phenotype do not necessarily have the same or similar annotation styles. Despite these findings, there are notable qualitative similarities identified between the aforementioned groups in the previous chapter. This leads to the question: how can one quantitatively identify similarities between annotation styles of different datasets? This thesis proposes an approach to identify subgroups between the different learned annotation styles, as described in Section 3.3. This chapter uses the proposed method to identify and confirm any similarity groupings between the annotation styles of the different trials. With the proposed analysis method, this chapter can build on the understanding of annotation styles and the factors that contribute to them. Additionally, in situations where datasets are small and aggre-

gation of datasets is necessary to obtain well-trained models, its especially important to know of existing relationships between annotation styles. If datasets have contradictory annotation styles, its extremely important to be aware and treat them independently. But in situations where annotation styles are fairly consistent between two datasets and data is scarce, pooling them together and treating them as one *group* can be beneficial. As such, in this chapter, we leverage the identified subgroups in a Group-Conditioned model to demonstrate the utility of the grouping method.

6.1 Subgroup Identification Results

This section explores and identifies relationships between the trial styles by exploiting relationships in the learned CIN parameters (parameters in the CIN layer) from the trained Trial-Conditioned model discussed in Section 5.1.2. Recall that there are a set of learned CIN parameters for each trial in the training set. The CIN parameters consist of a scale and shift parameter, each of size [1, channels, 1, 1, 1], and there are two CIN layers per block in the network, rendering analysis challenging. In order to effectively evaluate relationships between these high dimensional trial parameters, this chapter uses the proposed analysis using cosine similarity measures and vector norm values.

In the cosine similarity analysis, the metric is calculated as described in Section 3.3. This analysis allows the identification of similarities between the directions of the CIN parameters, with $+1$ indicating the vectors are in the same direction in the high dimensional space and -1 meaning they are in the opposite direction. This method allows one to gauge similarities between scales and shifts per layer, between each combination of trial pairs. For the vector norm analysis, the norm of the scale against the norm of the shift parameter is plotted in a scatter plot on a per-trial basis. This analysis results 14 scatter plots where each point is (scale, shift) for a different trial. This analysis, although only qualitative, is able to permit visualization of the different relationships considering magnitudes. For the purpose of brevity, only one scatter plot per section of the network

(encoder, center, decoder) is provided. The combined analysis of both cosine similarity and the scatter plots allow us to draw some conclusions with both direction and magnitude relationships between trials.

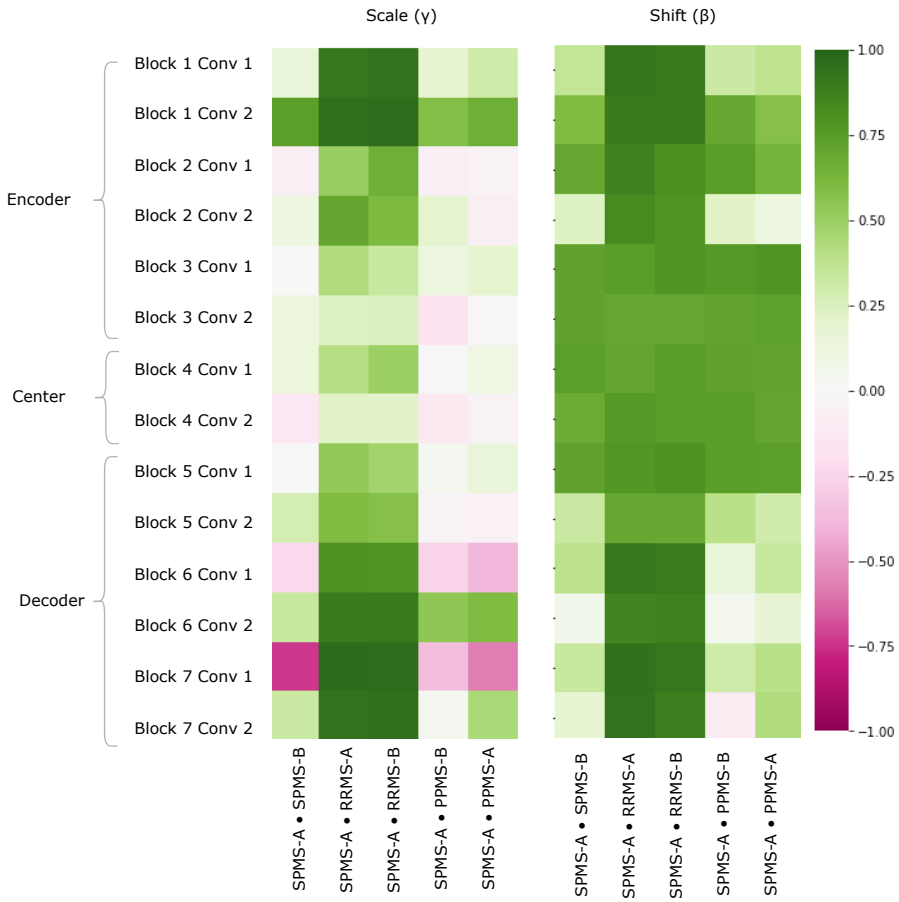


Figure 6.1: CIN parameter cosine similarity values between SPMS-A and all other trials for all CIN layers in the nnUNet.

Figure 6.1 provides a visualization of the (pairwise) cosine similarity analysis for SPMS-A with respect to all other trials. There are two fully green columns in 6.1, corresponding to the cosine similarity between SPMS-A and RRMS-A, and SPMS-A and RRMS-B. These green columns reflect that both RRMS-A and RRMS-B have similar trial-specific IN parameters as SPMS-A throughout the entire network. This means that the network is learning to apply similar affine transformations to RRMS-A and RRMS-B samples as learned for SPMS-A, suggesting similar learned styles. This results reveals the subgroup

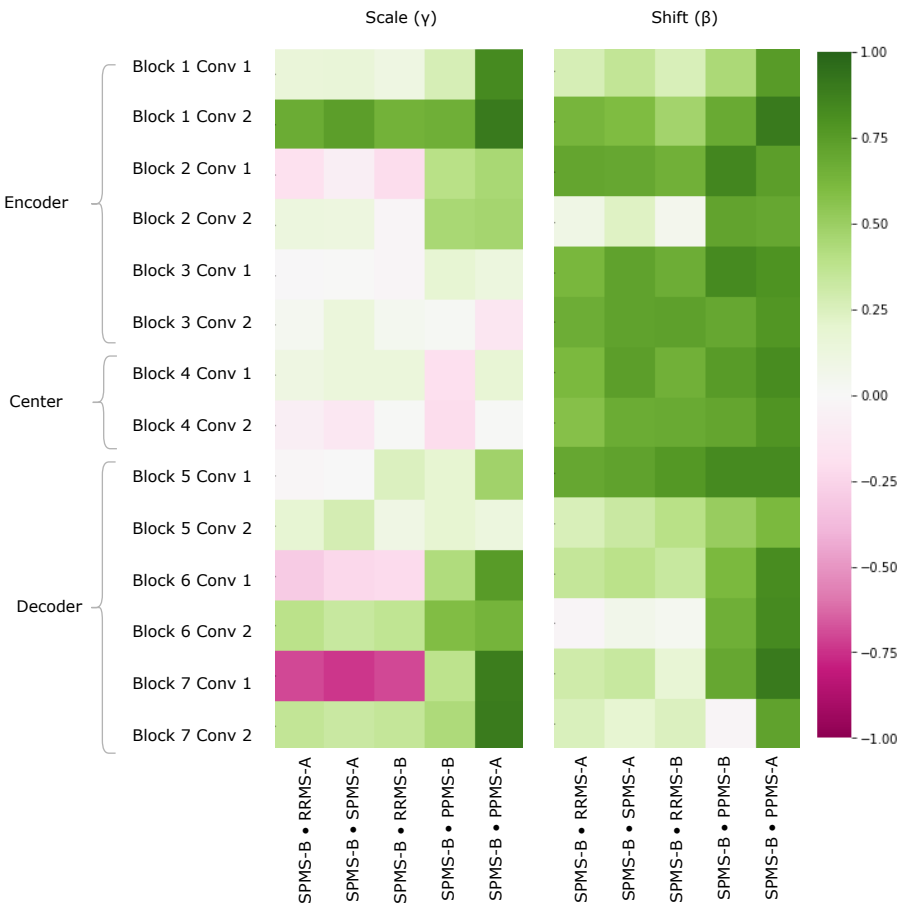


Figure 6.2: CIN parameter cosine similarity values between SPMS-B and all other trials for all CIN layers in the nnUNet.

trends noted in Section 5.1.2, where SPMS-A shows distinct similarity in annotation style with RRMS-A and RRMS-B. The same relationships are confirmed in all other trial comparisons, where two distinct style subgroups are discovered: (1) SPMS-B, PPMS-A, and PPMS-B, and (2) RRMS-A, RRMS-B, and SPMS-A. The main commonality between the groups is that the trials were labelled around approximately the same time (refer to Table 4.1), and as a result, with similar versions of protocols and semi-automated software. These results indicate that the label generation protocol is a main contributor to annotation bias between these particular trials. This relationship is further shown in the scatter plots in Figure 6.7 where the two groups form visible clusters. Figure 6.7 shows prox-

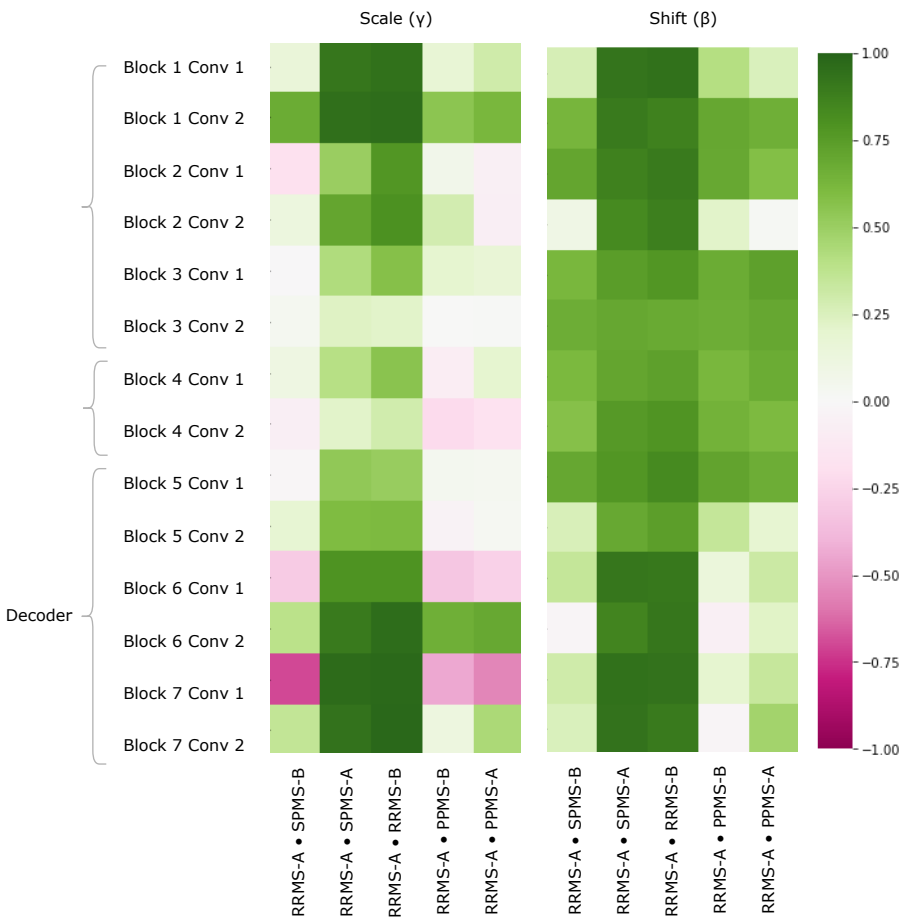


Figure 6.3: CIN parameter cosine similarity values between RRMS-A and all other trials for all CIN layers in the nnUNet.

imity between the SPMS-B/PPMS-B/PPMS-A group with some distance from the cluster formed by the RRMS-A/SPMS-A/RRMS-B group.

Other than the annotation style group identification, the cosine similarity analysis also uncovered other trends. Interestingly, the shifts are more similar across all trial comparisons than the scales, suggesting high discriminatory importance of the scale parameter in the CIN formula. The center layers were relatively very similar across many trials compared to the rest of the network layers, especially in the bias parameters. This could mean that there may not be a huge amount of distinction between trials while they are encoded into the latent space.

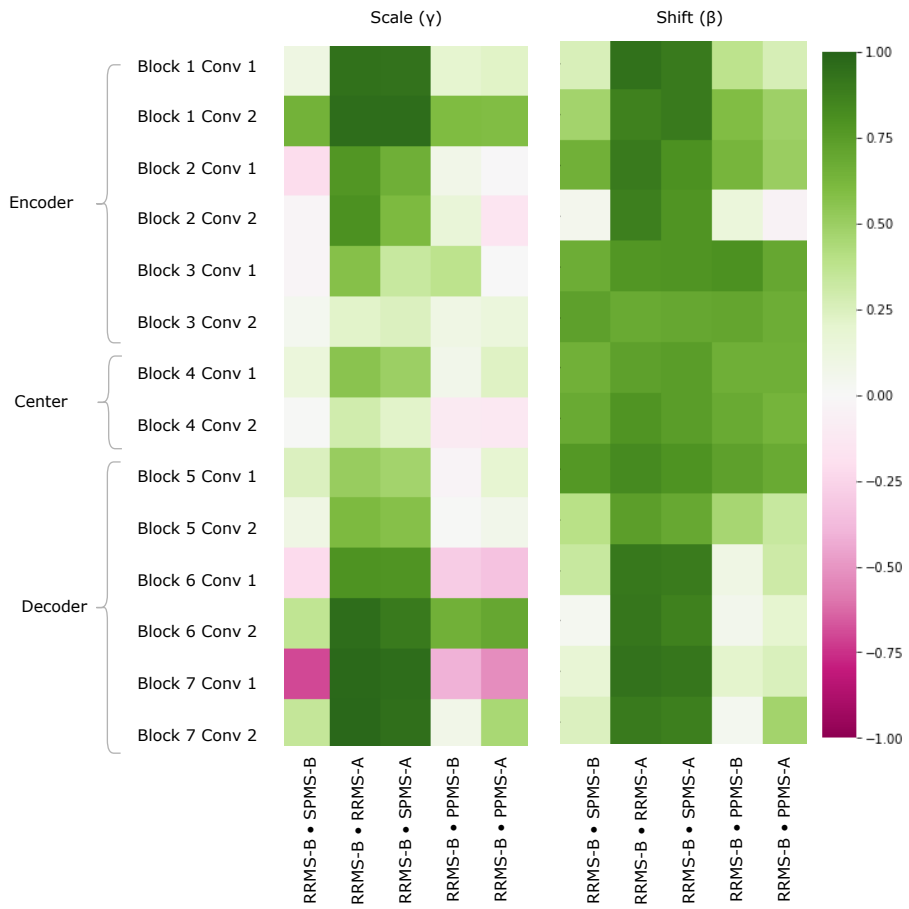


Figure 6.4: CIN parameter cosine similarity values between RRMS-B and all other trials for all CIN layers in the nnUNet.

Lastly, the PPMS-B trial results in Figure 6.6 show very little strong relation or opposition to most of the trials. Although the similarity with PPMS-A and SPMS-B is still noticeable, it is much less extreme than the similarity between just PPMS-A and SPMS-B. Figure 6.7 also shows that the PPMS-B point strays further from PPMS-A and SPMS-B. Figure 6.7 c) especially shows PPMS-B very astray from all other trial parameters. This trial's unique annotation could be due to a number of reasons, including unique changes or updates to the labelling or preprocessing pipeline, or a specific labelling protocol required for this patient cohort. These possible changes could all result PPMS-B's unique CIN parameters. Although similar to SPMS-B and PPMS-A, these trends lead us to believe the PPMS-B trial is its own group. Moving forward, we divide the data into three

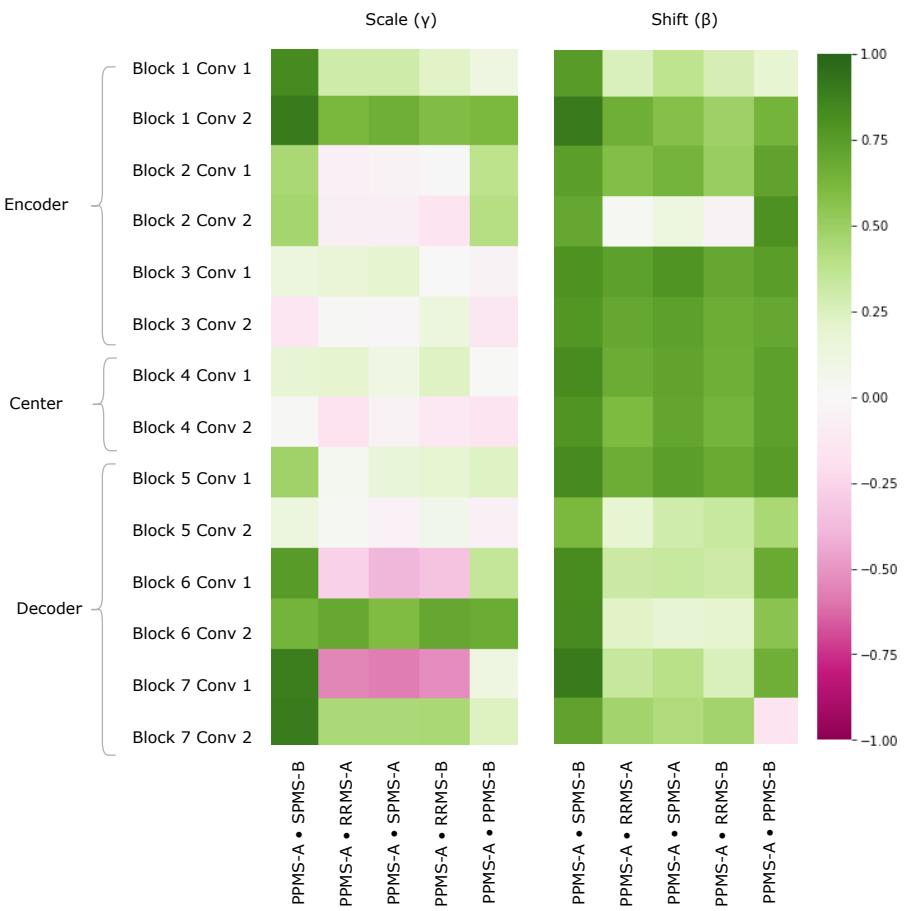


Figure 6.5: CIN parameter cosine similarity values between PPMS-A and all other trials for all CIN layers in the nnUNet.

groups: 1) RRMS-A, RRMS-B, SPMS-A, 2) SPMS-B, PPMS-A, and 3) PPMS-B. Upon discussion with our industrial partners, it was revealed that PPMS-B was initially processed using an entirely different pipeline and was reprocessed later in time. This finding emphasizes the importance of processing and labelling pipelines in data management and training DL networks. As most data comes from a variety sources and labelling protocols differ across research centers, this is likely a problem that affects many researchers. However, most researchers do not look into the data preparation differences and mostly focus on clinical or acquisition differences.

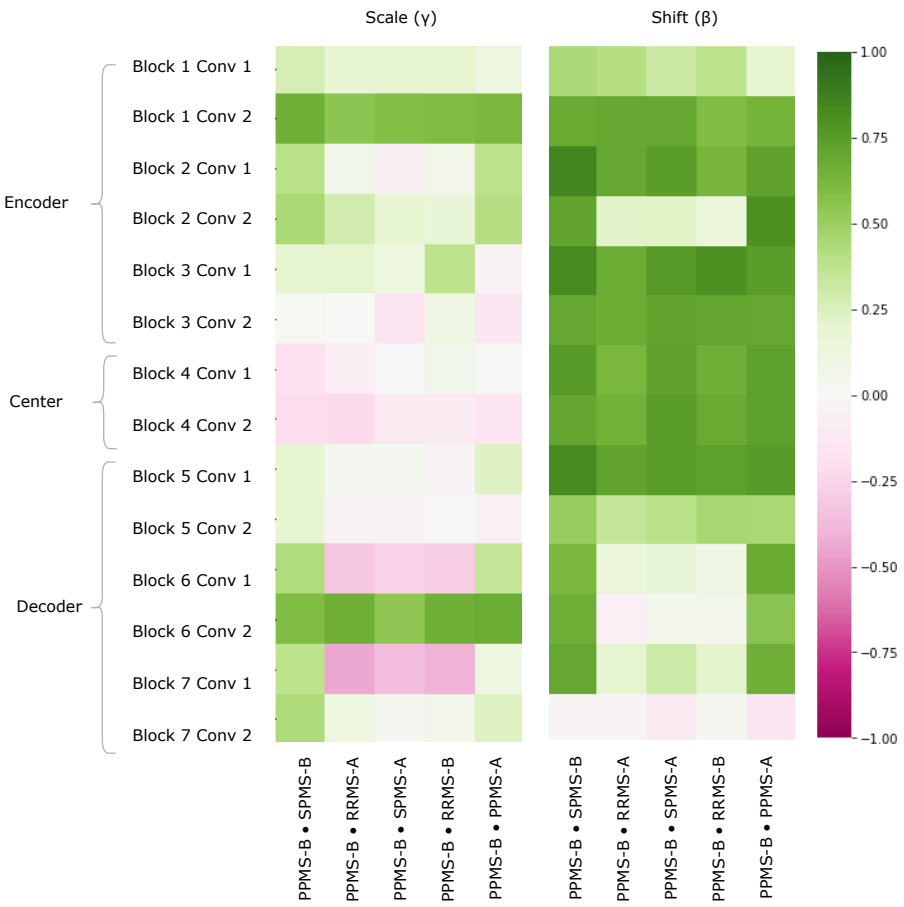
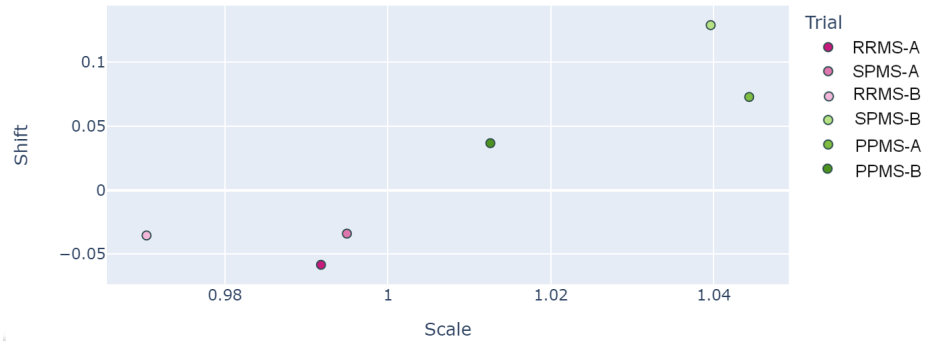
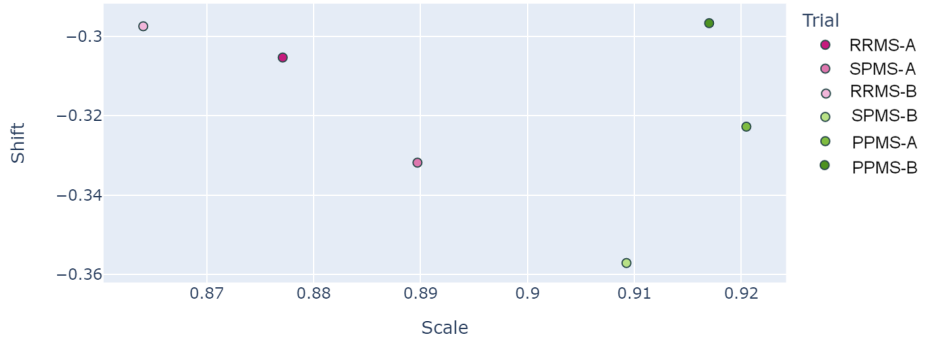


Figure 6.6: CIN parameter cosine similarity values between PPMS-B and all other trials for all CIN layers in the nnUNet.

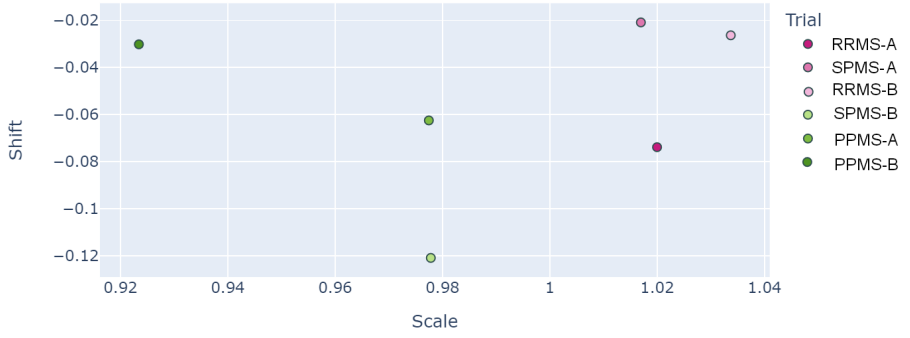
Fixating on acquisition differences or clinical factors may mislead some researchers in domain adaptation or data pooling tasks. For example, the identified groupings in this research contradict some of the clinical similarities between trials, specifically the disease subtype. While one might think that naively pooling trials of the same disease type would be appropriate, our analysis in this chapter, combined with results from Section 5.2, shows the importance of considering other factors in the labelling process. In the case of this study, the primary cause of the differences between trials was likely a difference in the labelling protocol and corresponding semi-automated labelling algorithm. The difference in labelling protocol could be due to the ambiguity in MS lesion boarders.



(a) Encoder scatter plot example (Block 1 Convolution 1)



(b) Center block scatter plot example (Block 4 Convolution 1).



(c) Decoder scatter plot example (Block 7 Convolution 1).

Figure 6.7: Scatter plots where each point shows the linear norm of scale and shift over all channels per trial for different layers. One example scatter plot is provided for each portion of the network (Encoder, Center, Decoder).

As previously noted in Section 2.3.2, the DAWM which can be found adjacent to both focal lesions and healthy tissue makes delineating lesion boarders or identifying smaller lesions somewhat subjective. The intensity distribution between DAWM and the focal

lesions overlap, thus forcing both annotation software and human raters to decide (sometimes unconsciously) on an arbitrary threshold to discretize what is in fact a continuous healthy-to-pathology transition [73]. Making assumptions about annotation styles due to some prior knowledge may mislead researchers and result in unfair analyses or comparisons between different datasets.

6.2 Leveraging Subgroups In Conditioned Models

The previous section outlined a strategy to explore annotation style relationships between trials by analysing similarities in the CIN parameters. The experimental results of the CIN parameter analysis indicated a strong relationship between the RRMS-A, RRMS-B, and SPMS-A trials. SPMS-B and PPMS-A also showed high CIN parameter similarities, and PPMS-B was identified as unique. We also showed earlier in the previous chapter that CIN could successfully model different annotation styles across an aggregated dataset allowing for more fair generalization evaluations. In this section, we leverage the group findings with a CIN-trained model, as well as several Naive-Pooled models in order to validate our identified groupings as well as to demonstrate their utility.

6.2.1 Experiment Details

This section conducts a series of experiments on the basis of the previously identified groupings. A single model is trained on each identified group (*Group-pooling*) without any conditioning, essentially a Naive-Pooled model. Another model is trained with conditioning on the identified trial groupings. This *Group-Conditioned* model was designed such that the trials within an identified group all share CIN parameters through the entire network during training. This results in one set of CIN parameters for SPMS-A, RRMS-A, and RRMS-B, and another set of parameters for SPMS-B and PPMS-A, and lastly a set for PPMS-B.

6.2.2 Results and Discussion

This section presents a comparison between the Group-Pooled, Group-Conditioned, and Trial-Conditioned models. The Group-Conditioned model and the different styles learned is also explored in detail.

Table 6.1: Performance on test sets for Group-Pooling models, Group-Conditioning model, and the Trial-Conditioned model.

(a) F1 performance on test sets.

	Model	Training Set	Testing Set (metric F1)					
			RRMS-A	RRMS-B	SPMS-A	SPMS-B	PPMS-A	PPMS-B
1	**Group-Pooling**	Group 1: RRMS-A, RRMS-B, SPMS-A	**0.792**	**0.794**	**0.792**	0.677	0.640	0.644
2	**Group-Pooling**	Group 2: SPMS-B, PPMS-A	0.705	0.704	0.693	0.735	0.714	0.696
3	**Group-Pooling**	Group 3: PPMS-B	0.669	0.671	0.663	0.669	0.646	0.742
4	**Trial-Conditioned**	**All**	0.787	0.789	0.788	**0.737**	0.709	0.744
5	**Group-Conditioned**	**All**	0.788	0.786	0.784	**0.737**	**0.715**	**0.746**

(b) PR-AUC performance on test sets.

	Model	Training Set	Testing Set (metric PR AUC)					
			RRMS-A	RRMS-B	SPMS-A	SPMS-B	PPMS-A	PPMS-B
1	**Group-Pooling**	Group 1: RRMS-A, RRMS-B, SPMS-A	**0.883**	**0.883**	**0.883**	0.747	0.696	0.703
2	**Group-Pooling**	Group 1: SPMS-B, PPMS-A	0.79	0.784	0.774	0.823	0.791	0.772
3	**Group-Pooling**	Group 3: PPMS-B	0.746	0.744	0.734	0.746	0.700	0.828
4	**Trial-Conditioned**	**All**	0.879	0.878	0.879	**0.826**	0.786	0.829
5	**Group-Conditioned**	**All**	0.881	0.876	0.876	**0.826**	**0.792**	**0.833**

The Group-Pooling models do not suffer from the same performance degradation previously demonstrated in the all-trial Naive-Pooling experiments. In Table 6.1, the resulting performance on the Group-Pooled model trained on SPMS-A, RRMS-A, RRMS-B group matches and slightly outperforms their Single-Trial baselines (see Table 5.1). The same is true for the SPMS-B and PPMS-A group-pooled model. These results confirm that the groups identified from the CIN parameter analysis do in fact have significant similarities between their annotation styles. The Group-Conditioned model also matches performance of the Trial-Conditioned model, further confirming the validity of the groups identified from the parameter analysis. These annotation style groups are especially beneficial

Table 6.2: Performance on test sets for the Group-Conditioned model using all annotation styles

(a) F1 performance on test sets.

	Conditioning Style	Testing Performance (F1)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	Group 1: RRMS-A, RRMS-B, SPMS-A	**0.788**	**0.786**	**0.784**	0.689	0.658	0.677
2	Group 2: SPMS-B, PPMS-A	0.709	0.705	0.694	**0.737**	**0.715**	0.701
3	Group 3:PPMS-B	0.693	0.695	0.684	0.7	0.663	**0.746**

(b) PR-AUC performance on test sets.

	Conditioning Style	Testing Performance (metric PR AUC)					
		RRMS-A	**RRMS-B**	**SPMS-A**	**SPMS-B**	**PPMS-A**	**PPMS-B**
1	Group 1: RRMS-A, RRMS-B, SPMS-A	**0.881**	**0.876**	**0.876**	0.769	0.728	0.751
2	Group 1: SPMS-B, PPMS-A	0.796	0.784	0.774	**0.826**	**0.792**	0.781
3	Group 3: PPMS-B	0.788	0.779	0.767	0.78	0.732	0.833

for cases where each dataset source is very small such that training parameters on each independent dataset source may not yield adequate performance in the desired annotation style. This can be demonstrated by the slight performance increase from the Single-Trial baselines for the RRMS-A, RRMS-B, and SPMS-A group. Since these datasets are already relatively large (training set of 234 patient volumes), the benefit of pooling them is quite small. But in situations where each individual dataset consists of very few samples, identifying these groups will likely lead to beneficial performance improvements by allowing the model to learn from more samples of each annotation style.

Table 6.2a also demonstrates that the Grouped-Conditioned model is successfully able to learn distinct styles even though several trials share CIN parameter sets. When using different annotation styles (conditioning on the "wrong" group), F1 scores suffer up to over 5% performance degradation and PR-AUC scores suffer approximately 10% degradation. This performance degradation further stresses the importance of understanding sources of annotation styles and their relationships. If researchers made assumptions

about the annotation styles from other auxiliary information, for example the disease subtype, they would make incorrect groupings and combine incompatible annotation styles. For example, in this case, SPMS-B and PPMS-A are not the same disease subtype, but they have compatible annotation styles. Contrarily, PPMS-B is the same subtype as PPMS-A, but as shown in Table 6.2, they have incompatible annotation styles resulting in the aforementioned performance degradation.

6.3 Summary

This chapter presented the results of a novel annotation style subgroup analysis method proposed in this thesis. The cosine similarity and vector norm analysis was able to uncover and confirm the main hypothesis in this thesis regarding the sources of annotation styles in our MS T2 segmentation dataset. The results confirmed that the annotation protocol, and associated semi-automated labelling software, is a major contributor to annotation style differences, rather than potential inter-rater variability, observer bias, or other factors. Given the abundance of semi-automated labelling software in modern medical segmentation tasks, this finding is of particular importance in the field. These results suggest that even datasets labelled using overlapping raters, similar training protocol for raters, and the same company software can still demonstrate annotation style shifts across datasets due to small or cumulative updates in the software.

In addition to providing insights on annotation styles, the subgroup analysis method also showed great utility in strategic dataset aggregation. The proposed analysis was able to identify datasets labelled with annotation styles compatible enough to be pooled together and treated as one. This is particularly of use in medical segmentation, where data is very limited, and conditioning each very small dataset may not be feasible. For some pathologies, even a sample size of 100 patients can be hard to come across. In these situations, every measure must be taken in order to ensure that all data is used in the most efficient manner. This chapter demonstrated that the proposed method for subgroup-

ing could allow for more effective and efficient use of the available data in aggregated datasets.

Chapter 7

Fine-tuning to Annotation Styles for Continual Learning

In the previous chapters, we showed the existence and the obstacle that is annotation styles in T2 lesion segmentation datasets. We presented a method to account for annotation style differences as well as to provide the user with a method of producing an output in all learned styles on any given sample. We also identified trials that were annotated with similar styles, likely due to the semi-automated labelling software used in generating the "ground truths". However, we did not address what to do when new data comes in with labels in an unseen annotation styles. In order to implement DL algorithms in the clinic, we need to ensure that they are adaptable to the evolving needs and pathology definitions used in any given clinic. Referring back to our Hospital A and B example from the introduction, consider a situation where a single trained model needs to be deployed in these hospitals for regular use. A model trained in a research lab or private company for these hospitals likely may not have seen data labelled with the style from Hospital A or Hospital B, and therefore will not be able to produce an output in said style required at either hospital. In order to provide these sites with the algorithm they need, this thesis proposes to fine-tune existing models.

Since a DL network cannot produce an output in a segmentation style it has not seen, its not fitting to deploy and evaluate said network in either Hospital A or Hospital B, because it will not be able to perform as needed. By fine-tuning to a select few labelled samples from the target hospital (or environment), we can expose the network to the desired annotation style and give it an opportunity to learn that style. This reduces the manual labor required to deploy an algorithm tailored to each center's needs. Without fine-tuning, these hospitals would have to generate fully labelled, large datasets in order to train their own model from scratch to produce their desired style. By fine-tuning only the CIN layers of the network, we reduce the risk of overfitting while only using a few labelled samples in the desired style. These CIN layers have proved important for style modelling in the experiments presented in this thesis, therefore this thesis proposes that fine-tuning them will result in quick adaptation to new annotation styles necessary for lifelong learning.

7.1 Experiment Details

For this series of experiments, a model is pre-trained on 4 *source* trails: RRMS-A, RRMS-B, SPMS-A, and PPMS-A. The *target* trial to fine-tune on is RRMS-C. These trials were selected to emulate a situation where the new, target trial has a different style from the previously seen trials. Given the results from the previous chapter, we know that the semi-automated software version (associated with the timing of collection) has a strong influence on the styles. As a result, a source group was selected from trials labelled with earlier versions of the semi-automated software, and a trial labelled with a later version of the semi-automated software was used as the target dataset.

In the following section, first the need for fine-tuning is demonstrated. The performance of a Trial-Conditioned and Naive-Pooled model trained only on the source datasets, without fine-tuning, are compared to the performance of an RRMS-C Single-Trial model. The RRMS-C Single-Trial model was trained on the entire RRMS-C training

set to establish a "ceiling" performance to represent how well a model can do if it has access to the entire 234 training samples of RRMS-C.

Next, fine-tuning from both a Trial-Conditioned model as well as a Naive-Pooled model is experimented with in order to demonstrate the utility of fine-tuning IN parameters in both the proposed model (Trial-Conditioned) of this thesis as well as the common practice model (Naive-Pooling). These models are fine-tuned on RRMS-C using the two different fine-tuning methods described in Section 3.4: Fine-Tuning from No Affine (initializing IN parameters at a scale of 1 and shift of 0) to Fine-Tuning from Learned Affine. When using the proposed approached of Fine-Tuning from Learned Affine with the CIN-trained model, one has the option of fine-tuning from each of the learned affine parameters of all source trials (ie fine-tuning from RRMS-A, RRMS-B, SPMS-A, and PPMS-A). In the following section, the results of fine-tuning from the learned affine parameters of the different trials are presented separately, and are referred to as Trial-Conditioned Fine-Tuned From $< TrialID >$. This is in contrast to using this strategy with Naive-Pooled models, where there is only one set of learned affine parameters to fine-tune from, which is simply referred to as Naive-Pooled Fine-Tuned From Learned Affine. For fine-tuning, 3 random non-overlapping sets of 5 RRMS-C training samples are selected. These 3 different sets are used to train 3 separate models to ensure that performance is not simply due irregularities or class balances in the selected 5 samples. The results for fine-tuned models are therefore presented as mean and variation of these 3 models trained on the 3 different sets. The Trial-Conditioned, Naive-Pooled, and Single-Trial models were all trained and hyperparameter tuned as described in Chapter 4. For the fine-tuned models, the hyperparameters were kept consisted with the source-trained model, as 5 samples is not enough to both fine-tune and hyperparameter tune on without risk of overfitting.

7.2 Results and Discussion

Here we detail the results reflecting the generalizability of models with and without fine-tuning. We also compare our proposed fine-tuning methods as previously described, and present some quantitative and qualitative results and discussion.

Table 7.1: Performance on the test set of RRMS-C for not fine-tuned, and fine-tuned models trained on various datasets. Note the RRMS-C Model is never fine-tuned, and it is a model trained on the entire RRMS-C training set.

(a) Performance on the test set of RRMS-C from various models with no fine-tuning.

	Model	Conditioning Style	Test PR-AUC	Test DICE
1	**Trial-Conditioned**	**RRMS-A**	0.730	0.662
2		**RRMS-B**	0.723	0.657
3		**SPMS-A**	0.726	0.659
4		**PPMS-A**	0.732	0.662
5	**Naive-Pooled**	-	0.729	0.664
6	**RRMS-C Model**	-	**0.793**	**0.712**

(b) Performance on the test set of RRMS-C from the Trial-Conditioned and Naive-Pooled models after fine-tuning to 5 labelled RRMS-C samples.

	Model	Fine-Tuned From	Test PR-AUC	Test DICE
1	**Trial-Conditioned**	**RRMS-A**	0.761 $\pm$0.00002	0.687 $\pm$0.00001
2		**RRMS-B**	0.762 $\pm$0.00002	0.687 $\pm$0.00001
3		**SPMS-A**	0.759 $\pm$0.00003	0.685 $\pm$0.00002
4		**PPMS-A**	0.763 $\pm$0.00004	0.688 $\pm$0.000004
5		**No Affine**	**0.765 $\pm$0.000005**	**0.692 $\pm$0.000004**
6	**Naive-Pooled**	**No Affine**	0.751 $\pm$0.000007	0.682 $\pm$0.00001
7		**Learned Affine**	0.763 $\pm$0.00004	0.690 $\pm$0.00003

We first demonstrate the need for fine-tuning to new annotation styles in Table 7.1a. In this table, the RRMS-C model is the Single-Trial, unconditioned model that is trained on the full RRMS-C dataset. Even though the Trial-Conditioned model and Naive-Pooled model have access to more than 4x the data compared to the RRMS-C model, they fall behind in performance by over 5% DICE and PR-AUC. No matter which conditioning style (CIN parameter sets) are used, the Trial-Conditioned model, which is usually somewhat

superior to the Naive-Pooled model, cannot produce results in the required segmentation style for RRMS-C. This clearly demonstrates that the styles of RRMS-A, RRMS-B, SPMS-A, and PPMS-A are not compatible enough with the style of RRMS-C. Without fine-tuning, neither the Naive-Pooled model nor the Trial-Conditioned model are able to produce the results that are consistent with the segmentation style of the ground-truth of RRMS-C. Contrary to some popular beliefs that if trained on enough diverse data an algorithm will be generalizable, here one can see that if annotation styles change in incoming datasets, algorithms will still not be able to generalize well without some exposure to the new annotation style.

Table 7.1b shows the performance of the previously presented Trial-Conditioned and Naive-Pooled models after fine-tuning on RRMS-C samples. Fine-tuning on only 5 samples resulted in approximately a 3% increase across the board. This performance is more comparable to the RRMS-C model performance. Prior to fine-tuning, Trial-Conditioned and Naive-Pooled models fell behind the RRMS-C model by approximately 6%. With fine-tuning from only a mere 5 labelled samples (compared to the 234 the RRMS-C model was trained on), that gap has closed by about half, demonstrating that fine-tuning only the IN affine parameters, whether it be from Learned Affine or from No Affine, shows valuable benefit. For instance, compare row 5 on Table 7.1a to row 7 in Table 7.1b, and note the performance of the Naive-Pooled model had improved by over 3% after being exposed to just 5 random labelled samples of RRMS-C.

For the Trial-Conditioned model, fine-tuning from any existing CIN parameters (From Learned Affine) lead to very similar performance as fine-tuning from No Affine, as shown by rows 1-4 on Table 7.1a and rows 1-5 on Table 7.1b. Fine-tuning from the learned SPMS-A CIN parameters lead to the lowest performance, but still by a very small margin. In the Naive-Pooled model, fine-tuning from No Affine is only $< 1\%$ worse than fine-tuning from Learned Affine parameters. Most-fine-tuning approaches will initialize from a pre-trained value (as done in the Fine-Tuning from Learned Affine approach), but for adapting to new annotation styles with CIN or IN layers, initializing the affine parameters

from No Affine (scale of 1 and shift of 0) is not particularly worse or better than initializing from Learned Affine values from a performance standpoint. Overall, fine-tuning the IN parameters of a network to a few labelled samples of a new dataset with a new style is shown to be successful and improve results compared to no fine-tuning.

Although these two approaches to fine-tuning the CIN parameters may not lead to many differences in performance metrics, there are two main considerations that need to be discussed: 1) the preservation of performance on source trials for continual learning, and 2) the potential differences in style from using the Fine-Tuned from Learned Affine method or the Fine-Tuned from No Affine method.

7.2.1 Performance Degradation on Source Trials After Fine-Tuning

Table 7.2: F1 performance on the source trial test sets of the Naive-Pooled model and the Trial-Conditioned model, before or after fine-tuning. Note that due to the implementation of the Trial-Conditioned model, performance on source trials is inherently maintained and unmodified, so the fine-tuned status is shown as null.

	Model	Fine-Tuned Status	F1 Performance on Original Trial Test Sets			
			RRMS-A	RRMS-B	SPMS-A	PPMS-A
1	**Trial-Conditioned**	**-**	**0.786**	**0.789**	**0.786**	**0.745**
2	**Naive-Pooled**	**NONE**	**0.788**	**0.789**	**0.788**	**0.744**
3		**From No Affine**	0.710 ±0.00009	0.699 ±0.0001	0.699 ±0.00009	0.723 ±0.000009
4		**From Learned Affine**	0.768 ±0.00008	0.759 ±0.0002	0.758 ±0.0001	0.730 ±0.00002

Ideally, a single model would be implemented in clinics or research centers that can work well on all source *and* target trials. However, if you fine-tune the Naive-Pooled model's IN parameters to a new annotation style, you must keep this new model separate, as the fine-tuned model no longer works well on the previously learned trials, as shown by the performance degradation in rows 2-4 in Table 7.2. This is problematic if the previously trials are resumed, which happens often as in the case of MAESTRO3 (SPMS-A) and DEFINE_ENDORSE (RRMS-B), which were trials that were resumed with some of the same patient cohort several years later. The annotation style for the old trials may need to be used again if the trials are ever resumed, in order to keep annotations consis-

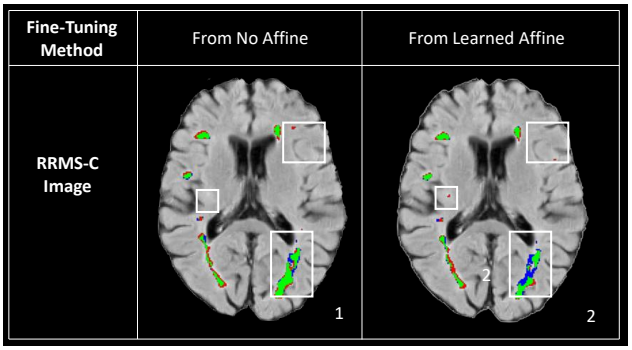
tent. This is also difficult for lifelong or continual learning as one would require a new fine-tuned model for every new trial coming in, as well as the original model for all other previously learned source trials. Instead of keeping all these models separate, this thesis recommends that it is best to simply modify the code for an existing model and add a *new* set of IN parameters to be trained in the fine-tuning stage. This essentially transitions a Naive-Pooled model to the Trial-Conditioning strategy, where each "dataset" has a different set of IN parameters, instead of all datasets sharing only 1 set. For the fine-tuned Naive-Pooled model, this would appear as a model having one set of IN parameters for the pool of source datasets, and one set of IN parameters fine-tuned to the target dataset. This is in contrast to having entirely separate Naive-Pooled models, pre- and post-fine-tuned to the target dataset, which is not practical in clinical implementations. Having new IN parameters for new target trials would prevent performance degradation on the source trials. In the Trial-Conditioned model and training strategy, all new trials possess their own CIN parameters, regardless of whether they were fine-tuned from No Affine or from Learned Affine. This allows Trial-Conditioned models to maintain performance on old trials, while still performing well on new trials, all within one model. This makes implementation in clinics and lifelong learning much easier.

7.2.2 Variations in Learned Annotation Styles

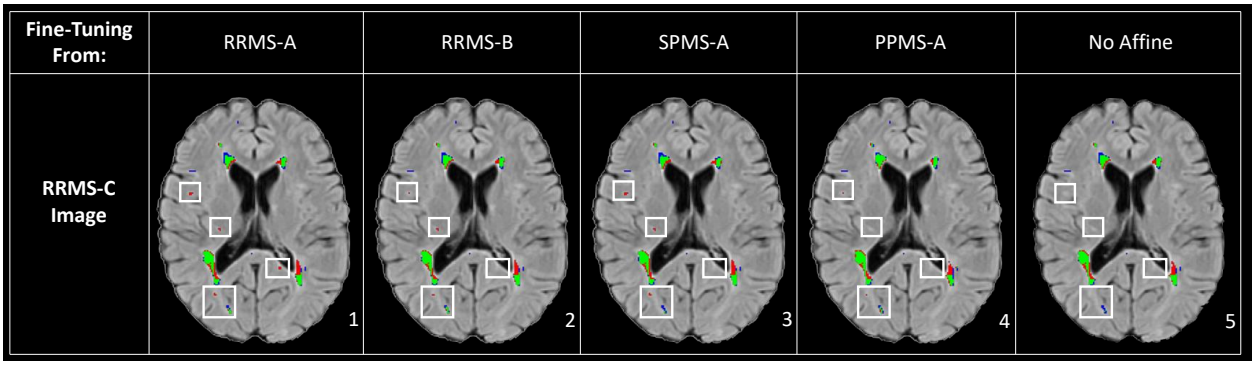
The second consideration regarding fine-tuning approaches is the potential difference in the style learned when the IN style parameters are learned from No Affine compared to Learned Affine values. When training on the source dataset(s), the IN affine parameters converge to the optimum for the respective trial (or the pooled dataset, in the Naive-Pooled model). When a new trial comes in for fine-tuning, and you initialize the IN parameters from previously learned IN parameters (as done in the fine-tuned from Learned Affine method), it forces the model to converge on new optimal affine parameters that are relative to the previously learned affine transform. Specifically, when a model is trained from scratch with IN or CIN layers, these layers are initialized with a scale of 1 and a shift

of zero, which essentially applies no affine transform. During training, the model then learns to apply any affine as needed. If, during fine-tuning, the model instead is forced to learn to modify an existing affine transform, it may converge on a different optimum than if it were learning to apply a *new* affine transform. By fine-tuning from existing IN parameters, you are changing the learning strategy and are instead forcing the algorithm to learn to change an already-learned affine. This could possibly bias the DL algorithm and cause it to learn a slightly different annotation style for the new target trial. Although fine-tuning from learned parameters is essentially the basis of fine-tuning any DL model, for style-specific parameters, this could have some potentially negative effects.

The effects of this are shown in Figure 7.1. There are noticeable differences in the styles learned by fine-tuning from No Affine compared to fine-tuning from Learned Affine IN parameters. These differences are not consistent or easily explainable either. Although the sample from the Naive-Pooled model shows significant under segmentation on one lesion when fine-tuned from Learned Affine parameters (Figure 7.1a box 2), under segmentation was not actually consistent across all the samples. Also note that there is a small false positive (all-red) lesion segmented in the brain shown in box 2). So despite the Fine-Tuned from Learned Affine model tending to under-segment the lesion in the lower right of the brain, it does have an over-detection element as well. The Fine-Tuned from No Affine model also detected a false-positive lesion in the top right that is missed by the Fine-Tuned from Learned Affine model. Recall that these types of differences were not apparent in the global metrics. DICE and PR-AUC are all global metrics and are unable to accurately present these small but definite changes in style. Furthermore, differences in segmentation style were also noticed in the Trial-Conditioned model (Figure 7.1b), and were similarly complex. A false-positive lesion is detected near the ventricle on the bottom right in the Fine-Tuned from RRMS-A model (box 1), and not in any other model. There are two other false-positive lesions on the left that are detected by the Fine-Tuned from RRMS-A, RRMS-B, and SPMS-A models but not in the remaining models. There is also a lesion in the bottom left that has different degrees of under-segmentation



(a) Example of an RRMS-C sample labelled with a Naive-Pooled model fine-tuned from No Affine or from Learned Affine instance normalization values.



(b) Example of an RRMS-C sample labelled with a Trial-Conditioned model fine-tuned from No Affine or from Learned Affine instance normalization values.

Figure 7.1: Examples showing how annotation styles differ between models with instance normalization parameters fine-tuned from No Affine and fine-tuned from Learned Affine. Green is True Positive, Red is False Positive, and Blue is False Negative. The white bounding boxes outline key points of differences between the segmentation maps.

between the 5 different fine-tuned models. These differences also reflect a subtle but definite change in the style learned depending on the type of fine-tuning approach taken, and depending on the set of learned affine parameters chosen for fine-tuning.

If a researcher knows for certain that a new annotation style coming in will be similar to a previously learned annotation style, it may instead be beneficial to fine-tune from the learned IN parameters of a similar trial. If this is not the case, and information regarding the styles of the source and target datasets is not known, it is best to use the Fine-Tune from No Affine strategy. This strategy does not suffer significantly in performance rel-

ative to the Fine-Tune from Learned Affine strategy, and may learn an annotation style that is less biased by the previously learned annotation style.

7.3 Summary

In this section, the importance of fine-tuning to trials with new annotation styles was demonstrated. This section proposed a training and fine-tuning strategy for lifelong learning to adapt to refining definitions of pathologies, changing clinical goals, and any other factors that affect annotation styles. Clear performance benefits were shown after using the proposed fine-tuning strategy on both the Trial-Conditioned model and the Naive-Pooled model. Although domain adaptation or few-shot meta-learning techniques could improve results in future work, fine-tuning is a much more computationally efficient and easy-to-implement method to quickly adapt a network to a new annotation style. This chapter showed the utility of implementing new IN parameters for each new style, initialized at a 1 scale and 0 shift, and tuning only the IN parameters to a few labelled samples. By adding a fresh set of IN parameters for each new, incoming trial and keeping them separate from all previously learned IN parameters, we can maintain performance on all old trials while still efficiently adapting to trials with unseen annotation styles as necessary.

Chapter 8

Conclusions

This thesis presented both an in depth analysis of annotation styles in focal pathology segmentation datasets, as well as a method to account for them in DL segmentation models. The primary contributions of this thesis were to bring attention to the issue of annotation styles and their impact on DL and generalization, as well as to propose a simple way to accommodate annotation style shifts across aggregated datasets in real-world practices. This thesis demonstrated the presence of annotation styles in MS segmentation datasets through generalizability assessments of Single-Trial models as well as through performance degradation of Naive-Pooled models. These experiments not only showed the existence of annotation styles, but also highlighted the importance of considering them when implementing DL algorithms in real clinics. Here we challenged notions of generalizability as well as ground truth in medical segmentation tasks. An in-depth evaluation of the proposed CIN method for annotation style modelling across 6 different clinical trials was conducted. The results from this analysis showed a clear benefit of the proposed conditioning framework over Naive-Pooled and Single-Trial models. The results also demonstrated the ability of CIN to model complex annotation styles. The CIN mechanism also provides practical benefits, by producing multiple segmentation masks, or "opinions" for any given sample. These multiple outputs can either be used to quantify uncertainty post-hoc, or can simply be used for providing more information to healthcare providers.

Furthermore, results from this thesis showed that although phenotype and observer bias may have an impact on some annotation styles, assumptions about annotation styles on the basis of phenotype or other patient demographics may not always be valid. The novel CIN parameter analysis method for identifying similar annotation styles across datasets also reinforced this finding. The results from this thesis strongly suggest that it is best to consider all aspects of the annotation process with great detail, from potential sources of bias in the human raters, to software updates in processing or labelling tools, and avoid simplifying assumptions. Lastly, this thesis proposed possible fine-tuning strategies for adapting existing networks to new data with unseen annotation styles. A flexible and sample-efficient fine-tuning strategy was presented that quickly adapts a network to new annotation styles for tailored implementation in clinics and research centres.

Throughout this entire thesis, the importance of considering annotation style shifts when evaluating an algorithm was highlighted; however, in most research, it is often not even considered. The results in this thesis showed that models that appear to "fail" to generalize may just be producing outputs in a different annotation style than that of the dataset ground-truth. As a result, other papers may be making incorrect conclusions about algorithm performance, generalizability, or adaptability. Although inter-rater bias or variability is a widely established phenomenon especially in medical tasks, other factors that contribute to variations in annotations are not accounted for. Even the concept of ground-truth is seldom challenged. As discussed in Chapter 2, many researchers simply assume that a ground-truth can and does exist, and variations between the raters is just a noisy reflection of the truth. This thesis instead poses that ground-truth is likely unattainable in many cases, especially from imaging data alone, and as a result, the entire model of ground-truth evaluations and generalizability should be approached with care.

The findings from this thesis can have several serious implications for the field of automated image segmentation overall. Although this thesis focused on focal pathology segmentation due to the high levels of ambiguity in the task, many other segmentation tasks that require advanced knowledge are likely similarly effected by annotation style

shifts. Furthermore, any segmentation task that uses semi-automated labelling assistance software, or uses strict protocols/instructions for raters can also become effected by annotation styles. As a result, it is important for all researchers to have as much knowledge on their dataset acquisition *and* labelling process as possible. Many public and private datasets keep their labelling information to themselves, as it can involve proprietary software; however, this thesis stresses the importance of data transparency. If the group collecting the data had certain goals, priorities, protocols, or software, these could all effect the annotations and therefore can effect how researchers should use the data, either in practice or in development. Both DICE and PR-AUC are only partially capable of quantifying annotation styles due to their global-scope. As a result, simple comparisons between a few sample ground-truths between datasets is not sufficient for identifying annotation style shifts. Unfortunately, this thesis had to rely on a combination of many different results including qualitative results, due to the nuance of annotation styles. Future work towards more sensitive metrics would improve knowledge on annotation styles as well as improve a researcher's ability to identify such problems early on. Future work could also be aimed at reducing stylistic differences between annotations from different sources; however, some factors contributing to annotation styles may be necessary for certain applications or tasks, such as study goal related factors, as outlined in the earlier Hospital A and Hospital B example. Furthermore, self-supervised or unsupervised learning techniques could avoid problems associated with annotation styles all together, but many of these approaches face challenges with the low number of samples and high dimensionality of medical images.

In conclusion, this thesis presents a novel, detailed analysis and discussion on systematic variations in annotations. This thesis shows that annotation styles a key factor, necessary to consider in the development, evaluation, and deployment of automated segmentation methods.

Bibliography

[1] AGHALARI, M., AGHAGOLZADEH, A., AND EZOJI, M. Brain tumor image segmentation via asymmetric/symmetric unet based on two-pathway-residual blocks. *Biomedical Signal Processing and Control 69* (2021), 102841.

[2] ALTAY, E. E., FISHER, E., JONES, S. E., HARA-CLEAVER, C., LEE, J.-C., AND RUDICK, R. A. Reliability of classifying multiple sclerosis disease activity using magnetic resonance imaging in a multiple sclerosis clinic. *JAMA neurology 70*, 3 (2013), 338–344.

[3] AMIRI, M., BROOKS, R., AND RIVAZ, H. Fine-tuning u-net for ultrasound image segmentation: different layers, different outcomes. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 67*, 12 (2020), 2510–2518.

[4] ASLANI, S., DAYAN, M., STORELLI, L., FILIPPI, M., MURINO, V., ROCCA, M. A., AND SONA, D. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. *NeuroImage 196* (2019), 1–15.

[5] AYALEW, Y. A., FANTE, K. A., AND MOHAMMED, M. A. Modified u-net for liver cancer segmentation from computed tomography images with a new class balancing method. *BMC Biomedical Engineering 3*, 1 (2021), 1–13.

[6] BATESON, M., KERVADEC, H., DOLZ, J., LOMBAERT, H., AND BEN AYED, I. Source-relaxed domain adaptation for image segmentation. In *International Confer-*

ence on Medical Image Computing and Computer-Assisted Intervention (2020), Springer, pp. 490–499.

[7] BIBERACHER, V., SCHMIDT, P., KESHAVAN, A., BOUCARD, C. C., RIGHART, R., SÄMANN, P., PREIBISCH, C., FRÖBEL, D., ALY, L., HEMMER, B., ET AL. Intra-and interscanner variability of magnetic resonance imaging based volumetry in multiple sclerosis. *Neuroimage 142* (2016), 188–197.

[8] BUI, M.-H., TRAN, T., TRAN, A., AND PHUNG, D. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems 34* (2021).

[9] CHEN, Y., ZHANG, H., WANG, Y., PENG, W., ZHANG, W., WU, Q. M. J., AND YANG, Y. D-bin: A generalized disentangling batch instance normalization for domain adaptation. *IEEE Transactions on Cybernetics* (2021), 1–13.

[10] CHOI, D., SHALLUE, C. J., NADO, Z., LEE, J., MADDISON, C. J., AND DAHL, G. E. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446* (2019).

[11] CHOTZOGLOU, E., AND KAINZ, B. Exploring the relationship between segmentation uncertainty, segmentation performance and inter-observer variability with probabilistic networks. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*. Springer, 2019, pp. 51–60.

[12] COHEN, J. P., HASHIR, M., BROOKS, R., AND BERTRAND, H. On the limits of cross-domain generalization in automated x-ray prediction. *ArXiv abs/2002.02497* (2020).

[13] CORONADO, I., GABR, R. E., AND NARAYANA, P. A. Deep learning segmentation of gadolinium-enhancing lesions in multiple sclerosis. *Multiple Sclerosis Journal 27*, 4 (2021), 519–527.

[14] DADAR, M., MAHMOUD, S., NARAYANAN, S., COLLINS, D. L., ARNOLD, D., AND MARANZANO, J. Diffusely abnormal white matter converts to t2 lesion volume in the absence of acute inflammation. *Brain* (2021).

[15] DAVIS, J., AND GOADRICH, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 233–240.

[16] FEDERATION, M. I. *Atlas of MS 3rd edition Part 1: Mapping Multiple Sclerosis Around the World*, 3 ed. MS International Federation, 2020.

[17] FILIPPI, M., PREZIOSA, P., LANGDON, D., LASSMANN, H., PAUL, F., ROVIRA, À., SCHOONHEIM, M. M., SOLARI, A., STANKOFF, B., AND ROCCA, M. A. Identifying progression in multiple sclerosis: new perspectives. *Annals of neurology 88*, 3 (2020), 438–452.

[18] FREEMAN, B., HAMMEL, N., PHENE, S., HUANG, A., ACKERMANN, R., KANZHELEVA, O., HUTSON, M., TAGGART, C., DUONG, Q., AND SAYRES, R. Iterative quality control strategies for expert medical image labeling. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (2021), vol. 9, pp. 60–71.

[19] GABR, R. E., CORONADO, I., ROBINSON, M., SUJIT, S. J., DATTA, S., SUN, X., ALLEN, W. J., LUBLIN, F. D., WOLINSKY, J. S., AND NARAYANA, P. A. Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study. *Multiple Sclerosis Journal 26*, 10 (2020), 1217–1226.

[20] GHAFOORIAN, M., MEHRTASH, A., KAPUR, T., KARSSEMEIJER, N., MARCHIORI, E., PESTEIE, M., GUTTMANN, C. R., LEEUW, F.-E. D., TEMPANY, C. M., GINNEKEN, B. V., ET AL. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention* (2017), Springer, pp. 516–524.

[21] GIRSHICK, R., DONAHUE, J., DARRELL, T., AND MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 580–587.

[22] GRAMMATIKOPOULOU, M., FLOUTY, E., KADKHODAMOHAMMADI, A., QUELLEC, G., CHOW, A., NEHME, J., LUENGO, I., AND STOYANOV, D. Cadis: Cataract dataset for surgical rgb-image segmentation. *Medical Image Analysis 71* (2021), 102053.

[23] GREENSPAN, H., VAN GINNEKEN, B., AND SUMMERS, R. M. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging 35*, 5 (2016), 1153–1159.

[24] HAUSER, S. L., AND CREE, B. A. Treatment of multiple sclerosis: a review. *The American journal of medicine 133*, 12 (2020), 1380–1390.

[25] HAVAEI, M., GUIZARD, N., CHAPADOS, N., AND BENGIO, Y. HeMIS: Hetero-modal image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2016), Springer, pp. 469–477.

[26] HELLER, N., DEAN, J., AND PAPANIKOLOPOULOS, N. Imperfect segmentation labels: How much do they matter? In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2018, pp. 112–120.

[27] HRÓBJARTSSON, A., THOMSEN, A. S. S., EMANUELSSON, F., TENDAL, B., HILDEN, J., BOUTRON, I., RAVAUD, P., AND BRORSON, S. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *Cmaj 185*, 4 (2013), E201–E211.

[28] HURWITZ, B. J. The diagnosis of multiple sclerosis and the clinical subtypes. *Annals of Indian Academy of Neurology 12*, 4 (2009), 226.

[29] ILSE, M., TOMCZAK, J. M., LOUIZOS, C., AND WELLING, M. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning* (2020), PMLR, pp. 322–348.

[30] ISENSEE, F., KICKINGEREDER, P., WICK, W., BENDSZUS, M., AND MAIER-HEIN, K. No new-net. In *International MICCAI Brainlesion Workshop* (2018), Springer, pp. 234–244.

[31] JACENKÓW, G., O'NEIL, A. Q., MOHR, B., AND TSAFTARIS, S. A. Inside: steering spatial attention with non-imaging information in cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), Springer, pp. 385–395.

[32] JADON, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (2020), IEEE, pp. 1–7.

[33] JAHANGARD, S., ZANGOOEI, M. H., AND SHAHEDI, M. U-net based architecture for an improved multiresolution segmentation in medical images. *arXiv preprint arXiv:2007.08238* (2020).

[34] JI, W., YU, S., WU, J., MA, K., BIAN, C., BI, Q., LI, J., LIU, H., CHENG, L., AND ZHENG, Y. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12341–12351.

[35] JOSKOWICZ, L., COHEN, D., CAPLAN, N., AND SOSNA, J. Inter-observer variability of manual contour delineation of structures in ct. *European radiology 29*, 3 (2019), 1391–1399.

[36] JUNGO, A., MEIER, R., ERMIS, E., BLATTI-MORENO, M., HERRMANN, E., WIEST, R., AND REYES, M. On the effect of inter-observer variability for a reliable es-

timation of uncertainty of medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), Springer, pp. 682–690.

[37] KAMNITSAS, K., BAUMGARTNER, C., LEDIG, C., NEWCOMBE, V., SIMPSON, J., KANE, A., MENON, D., NORI, A., CRIMINISI, A., RUECKERT, D., ET AL. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging* (2017), Springer, pp. 597–609.

[38] KARANI, N., CHAITANYA, K., BAUMGARTNER, C., AND KONUKOGLU, E. A lifelong learning approach to brain mr segmentation across scanners and protocols. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), Springer, pp. 476–484.

[39] KARIMI, D., DOU, H., WARFIELD, S. K., AND GHOLIPOUR, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Medical image analysis 65* (Oct. 2020), 101759. Publisher: NIH Public Access.

[40] KILJUNEN, T., AKRAM, S., NIEMELÄ, J., LÖYTTYNIEMI, E., SEPPÄLÄ, J., HEIKKILÄ, J., VUOLUKKA, K., KÄÄRIÄINEN, O.-S., HEIKKILÄ, V.-P., LEHTIÖ, K., ET AL. A deep learning-based automated ct segmentation of prostate cancer anatomy for radiation therapy planning-a retrospective multicenter study. *Diagnostics 10*, 11 (2020), 959.

[41] KIM, D., TSAI, Y.-H., SUH, Y., FARAKI, M., GARG, S., CHANDRAKER, M., AND HAN, B. Learning semantic segmentation from multiple datasets with label shifts. *arXiv preprint arXiv:2202.14030* (2022).

[42] KIM, Y., SOH, J. W., PARK, G. Y., AND CHO, N. I. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 3479–3489.

[43] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[44] KOMATSU, R., AND GONSALVES, T. Multi-cartoongan with conditional adaptive instance-layer normalization for conditional artistic face translation. *AI 3*, 1 (2022), 37–52.

[45] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems 25* (2012).

[46] LAFARGE, M. W., PLUIM, J. P., EPPENHOF, K. A., AND VETA, M. Learning domain-invariant representations of histological images. *Frontiers in medicine 6* (2019), 162.

[47] LANDMAN, B. A., ASMAN, A. J., SCOGGINS, A. G., BOGOVIC, J. A., STEIN, J. A., AND PRINCE, J. L. Foibles, follies, and fusion: Web-based collaboration for medical image labeling. *NeuroImage 59*, 1 (2012), 530–539.

[48] LECUN, Y., BENGIO, Y., AND HINTON, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[49] LI, W., MILLETARÌ, F., XU, D., RIEKE, N., HANCOX, J., ZHU, W., BAUST, M., CHENG, Y., OURSELIN, S., CARDOSO, M. J., ET AL. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging* (2019), Springer, pp. 133–141.

[50] LIAO, Z., HU, S., XIE, Y., AND XIA, Y. Modeling annotator preference and stochastic annotation error for medical image segmentation. *arXiv preprint arXiv:2111.13410* (2021).

[51] LIU, Y., DENG, J., TAO, J., CHU, T., DUAN, L., AND LI, W. Undoing the damage of label shift for cross-domain semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7042–7052.

[52] LUBLIN, F. D., REINGOLD, S. C., COHEN, J. A., CUTTER, G. R., SØRENSEN, P. S., THOMPSON, A. J., WOLINSKY, J. S., BALCER, L. J., BANWELL, B., BARKHOF, F., ET AL. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology 83*, 3 (2014), 278–286.

[53] MINAEE, S., BOYKOV, Y. Y., PORIKLI, F., PLAZA, A. J., KEHTARNAVAZ, N., AND TERZOPOULOS, D. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2021).

[54] MORADI, S., OGHLI, M. G., ALIZADEHASL, A., SHIRI, I., OVEISI, N., OVEISI, M., MALEKI, M., AND DHOOGE, J. Mfp-unet: A novel deep learning based approach for left ventricle segmentation in echocardiography. *Physica Medica 67* (2019), 58–69.

[55] MS SOCIETY OF CANADA. Multiple sclerosis society of canada asks canadians to #takeactionforms. `mssociety.ca/resources/news/article/multiple-sclerosis-society-of-canada-asks-canadians-to-takeactionforms`, 2020. Accessed: 2022-07-01.

[56] MURRAY, T. Clinical review-diagnosis and treatment of multiple sclerosis. *BMJ-British Medical Journal-International Edition 332*, 7540 (2006), 525–527.

[57] NAKKIRAN, P., KAPLUN, G., BANSAL, Y., YANG, T., BARAK, B., AND SUTSKEVER, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment 2021*, 12 (2021), 124003.

[58] NAM, H., AND KIM, H.-E. Batch-instance normalization for adaptively style-invariant neural networks. In *NeurIPS* (2018).

[59] NANA, A., RUTH, A. M., CHRISTINA, B., ROCHELLE, G., DOUGLAS, G. M., RON, W., PHILIPPE, F., JULIE, B., KAREN, T., AND KIM, R. Multiple sclerosis in canada 2011 to 2031: results of a microsimulation modelling study of epidemiological and economic impacts. *Health promotion and chronic disease prevention in Canada: research, policy and practice 37*, 2 (2017), 37.

[60] NICHYPORUK, B. Engineering deep learning systems for robust and accurate focal pathology segmentation and detection. Master's thesis, McGill University, 2022.

[61] NICHYPORUK*, B., CARDINELL*, J., SZETO, J., MEHTA, R., FALET, J.-P. R., ARNOLD, D. L., TSAFTARIS, S., AND ARBEL, T. Rethinking generalization: The impact of annotation style on medical image segmentation. *Under Review at: The Journal of Machine Learning for Biomedical Imaging* (2022).

[62] NICHYPORUK, B., CARDINELL, J., SZETO, J., MEHTA, R., TSAFTARIS, S., ARNOLD, D. L., AND ARBEL, T. Cohort bias adaptation in aggregated datasets for lesion segmentation. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health* (2021), Springer International Publishing, pp. 101–111.

[63] OAKDEN-RAYNER, L. Exploring large-scale public medical image datasets. *Academic radiology 27*, 1 (2020), 106–112.

[64] PHAM, D. L., XU, C., AND PRINCE, J. L. A survey of current methods in medical image segmentation. *Annual review of biomedical engineering 2*, 3 (2000), 315–337.

[65] RAJU, A., CHENG, C.-T., HUO, Y., CAI, J., HUANG, J., XIAO, J., LU, L., LIAO, C., AND HARRISON, A. P. Co-heterogeneous and adaptive segmentation from multi-source and multi-phase ct imaging data: a study on pathological liver and lesion segmentation. In *European Conference on Computer Vision* (2020), Springer, pp. 448–465.

[66] RECHT, B., ROELOFS, R., SCHMIDT, L., AND SHANKAR, V. Do imagenet classifiers generalize to imagenet? *ArXiv abs/1902.10811* (2019).

[67] RECHT, B., ROELOFS, R., SCHMIDT, L., AND SHANKAR, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning* (2019), PMLR, pp. 5389–5400.

[68] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[69] RUTA, D., MOTIIAN, S., FAIETA, B., LIN, Z. L., JIN, H., FILIPKOWSKI, A., GILBERT, A., AND COLLOMOSSE, J. P. Aladin: All layer adaptive instance normalization for fine-grained style similarity. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 11906–11915.

[70] SAITO, T., AND REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one 10*, 3 (2015), e0118432.

[71] SALEM, M., VALVERDE, S., CABEZAS, M., PARETO, D., OLIVER, A., SALVI, J., ROVIRA, À., AND LLADÓ, X. A fully convolutional neural network for new t2-w lesion detection in multiple sclerosis. *NeuroImage: Clinical 25* (2020), 102149.

[72] SCHWENKENBECHER, P., WURSTER, U., KONEN, F. F., GINGELE, S., SÜHS, K.-W., WATTJES, M. P., STANGEL, M., AND SKRIPULETZ, T. Impact of the mcdonald criteria 2017 on early diagnosis of relapsing-remitting multiple sclerosis. *Frontiers in neurology 10* (2019), 188.

[73] SEEWANN, A., VRENKEN, H., VAN DER VALK, P., BLEZER, E. L., KNOL, D. L., CASTELIJNS, J. A., POLMAN, C., POUWELS, P. J., BARKHOF, F., AND GEURTS, J. J. Diffusely abnormal white matter in chronic multiple sclerosis: imaging and histopathologic analysis. *Archives of neurology 66*, 5 (2009), 601–609.

[74] SEPAHVAND, N. M., ARNOLD, D. L., AND ARBEL, T. Cnn detection of new and enlarging multiple sclerosis lesions from longitudinal mri using subtraction images. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)* (2020), IEEE, pp. 127–130.

[75] SHANKAR, V., ROELOFS, R., MANIA, H., FANG, A., RECHT, B., AND SCHMIDT, L. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning* (2020), PMLR, pp. 8634–8644.

[76] SHARMA, N., AND AGGARWAL, L. M. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India 35*, 1 (2010), 3.

[77] SHARMA, S., SHARMA, S., AND ATHAIYA, A. Activation functions in neural networks. *International Journal of Engineering Applied Sciences and Technology 4*, 12 (2020), 310–316.

[78] SHEN, Y., AND GAO, M. Brain Tumor Segmentation on MRI with Missing Modalities. In *Information Processing in Medical Imaging* (Cham, 2019), A. C. S. Chung, J. C. Gee, P. A. Yushkevich, and S. Bao, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 417–428.

[79] SHINOHARA, R. T., OH, J., NAIR, G., CALABRESI, P. A., DAVATZIKOS, C., DOSHI, J., HENRY, R. G., KIM, G., LINN, K. A., PAPINUTTO, N., ET AL. Volumetric analysis from a harmonized multisite brain mri study of a single subject with multiple sclerosis. *American Journal of Neuroradiology 38*, 8 (2017), 1501–1509.

[80] SHWARTZMAN, O., GAZIT, H., SHELEF, I., AND RIKLIN-RAVIV, T. The worrisome impact of an inter-rater bias on neural network training. *arXiv preprint arXiv:1906.11872* (2019).

[81] SILVA, S., GUTMAN, B. A., ROMERO, E., THOMPSON, P. M., ALTMANN, A., AND LORENZI, M. Federated learning in distributed medical databases: Meta-analysis

of large-scale subcortical brain data. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)* (2019), IEEE, pp. 270–274.

[82] SIMPSON, A. L., ANTONELLI, M., BAKAS, S., BILELLO, M., FARAHANI, K., VAN GINNEKEN, B., KOPP-SCHNEIDER, A., LANDMAN, B. A., LITJENS, G., MENZE, B., ET AL. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).

[83] SOLOMON, A. J., NAISMITH, R. T., AND CROSS, A. H. Misdiagnosis of multiple sclerosis: Impact of the 2017 mcdonald criteria on clinical practice. *Neurology 92*, 1 (2019), 26–33.

[84] SOULAMI, K. B., KAABOUCH, N., SAIDI, M. N., AND TAMTAOUI, A. Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using unet model based-semantic segmentation. *Biomedical Signal Processing and Control 66* (2021), 102481.

[85] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*, 1 (2014), 1929–1958.

[86] SUN, B., AND SAENKO, K. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision* (2016), Springer, pp. 443–450.

[87] TAHA, A. A., AND HANBURY, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging 15*, 1 (2015), 1–28.

[88] TAJBAKHSH, N., SHIN, J. Y., GURUDU, S. R., HURST, R. T., KENDALL, C. B., GOTWAY, M. B., AND LIANG, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging 35*, 5 (2016), 1299–1312.

[89] TAKAHASHI, S., TAKAHASHI, M., KINOSHITA, M., MIYAKE, M., KAWAGUCHI, R., SHINOJIMA, N., MUKASA, A., SAITO, K., NAGANE, M., OTANI, R., ET AL. Fine-tuning approach for segmentation of gliomas in brain magnetic resonance images with a machine learning method to normalize image differences among facilities. *Cancers 13*, 6 (2021), 1415.

[90] TOMAR, D., LORTKIPANIDZE, M., VRAY, G., BOZORGTABAR, B., AND THIRAN, J.-P. Self-attentive spatial adaptive normalization for cross-modality domain adaptation. *IEEE Transactions on Medical Imaging 40*, 10 (2021), 2926–2938.

[91] VĂDINEANU, , PELT, D., DZYUBACHYK, O., AND BATENBURG, J. An analysis of the impact of annotation errors on the accuracy of deep learning for cell segmentation. In *Proceedings of the Conference on Medical Imaging with Deep Learning* (2021).

[92] VAN OPBROEK, A., IKRAM, M., VERNOOIJ, M., AND DE BRUIJNE, M. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging 34*, 5 (2014), 1018–1030.

[93] VINCENT, D., JONATHON, S., AND MANJUNATH, K. A learned representation for artistic style. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (2017), OpenReview.net.

[94] VINCENT, O., GROS, C., AND COHEN-ADAD, J. Impact of individual rater style on deep learning uncertainty in medical imaging segmentation. *arXiv preprint arXiv:2105.02197* (2021).

[95] VRENKEN, H., JENKINSON, M., PHAM, D. L., GUTTMANN, C. R., PARETO, D., PAARDEKOOPER, M., DE SITTER, A., ROCCA, M. A., WOTTSCHEL, V., CARDOSO, M. J., ET AL. Opportunities for understanding ms mechanisms and progression with mri using large-scale data sharing and artificial intelligence. *Neurology 97*, 21 (2021), 989–999.

[96] WANG, J., PEREZ, L., ET AL. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit 11* (2017), 1–8.

[97] WANG, R., CHAUDHARI, P., AND DAVATZIKOS, C. Embracing the disharmony in medical imaging: A simple and effective framework for domain adaptation. *Medical Image Analysis 76* (2022), 102309.

[98] WANG, X., YANG, X., DOU, H., LI, S., HENG, P.-A., AND NI, D. Joint segmentation and landmark localization of fetal femur in ultrasound volumes. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (2019), IEEE, pp. 1–5.

[99] WARFIELD, S., ZOU, K., AND WELLS, W. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging 23*, 7 (2004), 903–921.

[100] YAMASHITA, R., NISHIO, M., DO, R. K. G., AND TOGASHI, K. Convolutional neural networks: an overview and application in radiology. *Insights into imaging 9*, 4 (2018), 611–629.

[101] YAN, W., WANG, Y., GU, S., HUANG, L., YAN, F., XIA, L., AND TAO, Q. The domain shift problem of medical image segmentation and vendor-adaptation by unet-gan. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (Cham, 2019), D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Springer International Publishing, pp. 623–631.

[102] YIN, P., CAI, H., AND WU, Q. Df-net: Deep fusion network for multi-source vessel segmentation. *Information Fusion 78* (2022), 199–208.

[103] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B., AND VINYALS, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM 64*, 3 (2021), 107–115.

[104] ZHANG, L., TANNO, R., XU, M., JIN, C., JACOB, J., CICCARELLI, O., BARKHOF, F., AND ALEXANDER, D. C. Disentangling human error from the ground truth in segmentation of medical images. *arXiv preprint arXiv:2007.15963* (2020).

[105] ZHANG, Q., LIU, L., MA, K., ZHUO, C., AND ZHENG, Y. Cross-denoising network against corrupted labels in medical image segmentation with domain shift. *arXiv preprint arXiv:2006.10990* (2020).