DISS. ETH No. 28281

TACKLING DISTRIBUTION SHIFTS IN MACHINE LEARNING-BASED MEDICAL IMAGE ANALYSIS

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH ZURICH)

presented by

NEERAV KARANI

M. sc. ETH ZURICH

born on 31 October 1989

citizen of India

accepted on the recommendation of

Prof. Dr. Ender Konukoglu, examiner

Prof. Dr. Ben Glocker, co-examiner

Prof. Dr. Rama Chellappa, co-examiner

2022

ABSTRACT

Machine learning algorithms - in particular, those based on convolutional neural networks (CNNs) - have demonstrated remarkable promise in a number of medical image analysis tasks, even rivalling accuracies of human experts in some cases. A key requirement for good performance of these data-driven methods, however, is that their training and test samples must belong to the same probability distribution. This premise is often violated in medical imaging.

Training datasets often insufficiently represent image variations that may potentially occur at test time. For instance, if the training dataset is collected from a small number of clinics, it may overly emphasize the acquisition protocols of those clinics. If the trained algorithm is now used in a different hospital, the test images may be acquired with a scanner from a different vendor, and with a different acquisition protocol. Such test images, although potentially consisting of similar anatomical structures or lesions as in the training images, often differ from training images in terms of contrast and signal-to-noise ratio. This scenario, where the distribution of test samples differs from that of the training samples, is referred to as a distribution shift (DS). (We use the acronym DS to refer to both, the singular ’distribution shift’ and the plural ’distribution shifts’, and call on the reader to infer the form based on the context.) The aforementioned example is one of acquisition-related DS. Other DS such as those stemming from population differences (e.g. young v/s old) or disease effects (e.g. healthy v/s diseased) are also equally pertinent in medical imaging. CNN-based methods, that work very well on test samples from the training distribution, often exhibit remarkable performance degradation when faced with DS.

In this thesis, we develop three approaches with increasing generality to improve robustness of CNN-based medical image analysis methods in the presence of acquisition-related DS.

First, we develop a transfer learning approach for brain MRI segmentation across scanners and imaging protocols. A segmentation CNN’s batch normalization parameters are treated as distribution-specific, and tuned for each new test distribution using a small number of labelled images. The rest of the parameters are considered as distribution-agnostic; these are trained using the training dataset, and kept unchanged for test images.

Second, we consider a setting where (a) no labelled test images are available for transfer learning, (b) even unlabelled images from the test distribution are unavailable at training time, and (c) the training dataset cannot be transported to the test site. In this setting, we develop a test-time adaptation (TTA) method for medical image segmentation. A shallow normalization module of the segmentation CNN is adapted specifically for each test image. The adaptation relies on an implicit prior in the output space, which is modeled using a denoising autoencoder. Such a prior model can be considered as

a helper module, which encourages predicted segmentations that are similar to those seen during training.

Third, we note that if the helper module for TTA is itself modeled using a CNN, it is also likely to suffer from the DS problem. That is, the outputs of the helper module may be unreliable when the distribution of its test inputs differs from that of its training inputs. To this end, we employ field-of-experts (FoEs) to model the distribution in the output space of the adaptable normalization module. FoEs model high-dimensional probability distributions as a product of multiple low-dimensional distributions, and have improved robustness to DS. We use the task CNN's convolutional filters as the experts in the FoE model, and extend the model with additional experts as projections onto principal components of the task CNN's last layer features. This method is task-agnostic as the helper model is generically defined in the space of normalized images, rather than in a task-specific output space.

ZUSSAMMENFASSUNG

Algorithmen des maschinellen Lernens - insbesondere solche, die auf Faltungsneuronalen Netzen (CNN) basieren - haben sich bei einer Reihe von Aufgaben der medizinischen Bildanalyse als bemerkenswert vielversprechend erwiesen und können es in einigen Fällen sogar mit der Genauigkeit menschlicher Experten aufnehmen. Eine wichtige Voraussetzung für eine gute Leistung dieser datengesteuerten Methoden ist jedoch, dass ihre Trainings- und Teststichproben der gleichen Wahrscheinlichkeitsverteilung angehören müssen. Diese Voraussetzung wird in der medizinischen Bildgebung häufig verletzt.

Trainingsdatensätze repräsentieren oft nur unzureichend Bildvariationen, die zum Zeitpunkt der Prüfung auftreten können. Wenn der Trainingsdatensatz beispielsweise in einer kleinen Anzahl von Kliniken gesammelt wurde, kann er die Aufnahmeprotokolle dieser Kliniken übermäßig betonen. Wenn der trainierte Algorithmus nun in einem anderen Krankenhaus verwendet wird, werden die Testbilder möglicherweise mit einem Scanner eines anderen Herstellers und mit einem anderen Aufnahmeprotokoll aufgenommen. Solche Testbilder enthalten zwar möglicherweise ähnliche anatomische Strukturen oder Läsionen wie die Trainingsbilder, unterscheiden sich aber häufig in Bezug auf Kontrast und Signal-Rausch-Verhältnis von den Trainingsbildern. Dieses Szenario, bei dem sich die Verteilung der Testproben von derjenigen der Trainingsproben unterscheidet, wird als distribution shift (DS). (Wir verwenden das Akronym DS sowohl für den Singular "Verteilungsverschiebungäls auch für den Plural "Verteilungsverschiebungenünd fordern den Leser auf, die Form aus dem Kontext abzuleiten.) Das oben genannte Beispiel ist eines der akquisitationbezogenen DS. Andere DS, wie z. B. solche, die sich aus Bevölkerungsunterschieden (z. B. jung vs. alt) oder Krankheitsauswirkungen (z. B. gesund vs. krank) ergeben, sind in der medizinischen Bildgebung ebenfalls von Bedeutung. CNN-basierte Methoden, die bei Testproben aus der Trainingsverteilung sehr gut funktionieren, zeigen oft eine bemerkenswerte Leistungsverschlechterung, wenn sie mit DS konfrontiert werden.

In dieser Arbeit entwickeln wir drei Ansätze mit zunehmender Allgemeingültigkeit, zur Verbesserung der Robustheit von CNN-basierten medizinischen Bildanalysemethoden in Anwesenheit von akquisitationbezogenen DS.

Zunächst entwickeln wir einen Transfer-Learning-Ansatz für die MRT-Segmentierung des Gehirns über verschiedene Scanner und Bildgebungsprotokolle hinweg. Die Parameter für die Stapelnormalisierung eines Segmentierungs-CNN werden als verteilungsspezifisch behandelt und für jede neue Testverteilung anhand einer kleinen Anzahl von markierten Bildern abgestimmt. Die übrigen Parameter werden als verteilungsunabhängig betrachtet. Sie werden anhand des Trainingsdatensatzes trainiert und für Testbilder unverändert beibehalten.

Zweitens betrachten wir eine Situation, in der (a) keine beschrifteten Testbilder für das Transferlernen zur Verfügung stehen, (b) selbst unbeschriftete Bilder aus der Testverteilung zum Trainingszeitpunkt nicht verfügbar sind und (c) der Trainingsdatensatz nicht zum Testort transportiert werden kann. Vor diesem Hintergrund entwickeln wir ein Test-Zeit-Anpassungsverfahren (TTA) für die Segmentierung medizinischer Bilder. Ein flaches Normalisierungsmodul des Segmentierungs-CNN wird speziell für jedes Testbild angepasst. Die Anpassung stützt sich auf einen impliziten Prior im Ausgaberaum, der mit einem Denoising-Autoencoder modelliert wird. Ein solches Prior-Modell kann als Hilfsmodul betrachtet werden, das vorausgesagte Segmentierungen fördert, die den beim Training gesehenen ähnlich sind.

Drittens ist anzumerken, dass das Hilfsmodul für TTA, wenn es selbst mit einem CNN modelliert wird, wahrscheinlich auch unter dem DS-Problem leiden wird. Das heißt, die Ausgaben des Hilfsmoduls können unzuverlässig sein, wenn die Verteilung der Testeingaben von der Verteilung der Trainingseingaben abweicht. Zu diesem Zweck verwenden wir Field-of-Experts (FoEs), um die Verteilung im Ausgaberaum des anpassungsfähigen Normalisierungsmoduls zu modellieren. FoEs modellieren hochdimensionale Wahrscheinlichkeitsverteilungen als ein Produkt mehrerer niedrigdimensionaler Verteilungen und haben eine verbesserte Robustheit gegenüber DS. Wir verwenden die Faltungsfilter des Aufgaben-CNN als Experten im FoE-Modell und erweitern das Modell mit zusätzlichen Experten als Projektionen auf die Hauptkomponenten der Merkmale der letzten Schicht des Aufgaben-CNN. Diese Methode ist aufgabenunabhängig, da das Hilfsmodell generisch im Raum der normalisierten Bilder und nicht in einem aufgabenspezifischen Ausgaberaum definiert ist.

ACKNOWLEDGEMENTS

I thank Ender for giving me the opportunity to be a part of his outstanding group. I could not have asked for a better PhD advisor. The things that he has taught me in the last five years are too many to list. Most of all, he has shown by example, the importance of doing theoretically sound as well as impactful research, of upholding scientific integrity, and of keeping a fine balance of zooming in to the details and zooming out to the big picture.

I thank Christine, Christian, Krishna, Kerem, Xiaoran, Gustav, Anna, Ertunc, Kyriakos, Georg, Jonatan, Meva, Sara and Alex for creating a friendly and supportive work environment. Each of them brought a peculiar flair to the group, and my research journey has been heavily enriched by intriguing discussions with each of them on many occasions. I will fondly recall lunch conversations with Kerem and Anna. I apologize to all my table-football partners, for they were invariably on the losing side.

In the larger CVL family, I had the pleasure of many a good conversation with Christos, Danda, Samarth, Vaishakh, Bhaskar, Firat and Arun. I had the opportunity to interact with and to supervise several brilliant master's students for their semester or master theses. From Lin, Silvan, Yigit, Nicolas, Pol, Carina, Simin, Hande, Max, Zhexin, Aris and Dinos, I have learned as much, if not more than what I could teach them.

I thank Harry for resolving my IT issues, and Andreas and Krishna for helping me out with my numerous GPU-related questions. It is safe to say that their contributions hastened my research by several weeks, if not months. I am also deeply in awe with the efficiency of the CVL secretaries - Christina, Christine, Christina and Kristine. Their immense work often occurs in the background but was surely responsible for concealing my inabilities in handling administrative things.

Without Kruti, none of this would have been possible. She kept me grounded in my highs and lifted me from numerous setbacks. I thank her for always being there for me, for showing me that a world exists outside ETF E112 and for bearing with my mood swings. Last, but foremost, I thank my parents for their unwavering support and for teaching me to be content without being complacent.

Contents

1 Introduction **1**
1.1 Motivation for Algorithmic Analysis of Medical Images 1
1.1.1 Common Tasks in Medical Image Analysis 3
1.2 Methods for Medical Image Analysis 4
1.2.1 Hand-crafted Models 4
1.2.2 Data-Driven Models 5
1.3 Robustness to Distribution Shifts 7
1.3.1 The Distribution Shift Problem 7
1.3.2 Distribution Shifts in Medical Imaging 7
1.4 Contributions 10
1.5 Layout of the Thesis 13

2 Literature review **14**
2.1 Notation 14
2.2 Machine Learning Settings to Tackle Distribution Shifts 15
2.2.1 Supervised Learning 16
2.2.2 Transfer Learning 17
2.2.3 Unsupervised Domain Adaptation 19
2.2.4 Domain Generalization 21
2.2.5 Test-Time Adaptation 23
2.2.6 Source-Free Domain Adaptation 25
2.2.7 Post-Processing Methods 25
2.2.8 Unsupervised Learning 26
2.3 Problems Related to Distribution Shift Robustness 26
2.3.1 Out-of-distribution Detection 26
2.3.2 Model Performance Prediction 27
2.3.3 Adversarial Robustness 28

3 Datasets **29**
3.1 Datasets for Image segmentation 29
3.1.1 Reasons for DS in the used Datasets 30
3.1.2 Prostate 32
3.1.3 Heart 33
3.1.4 Spine 34
3.1.5 Brain (Healthy structure segmentation) 34
3.1.6 Brain (Cerebral White Matter Hyper-intensities) 35
3.2 Datasets for Image registration 38

4 A Transfer Learning Approach for Robust Medical Image Segmentation 39
4.1 Introduction . . . 39
4.2 Background . . . 40
4.2.1 Batch Normalization . . . 40
4.2.2 Common CNN Architectures for Image Segmentation . . . 41
4.3 Method . . . 42
4.3.1 Splitting Parameters into Distribution-Agnostic and Distribution-Specific . . . 42
4.3.2 Supervised Learning on Training Distributions . . . 42
4.3.3 Estimating the Closest Training Distribution to the Given Test Distribution . . . 42
4.3.4 Transfer Learning on the Given Test Distribution . . . 42
4.3.5 Favourable Properties for Lifelong Learning . . . 43
4.4 Experiments and Results . . . 43
4.4.1 Datasets . . . 43
4.4.2 Pre-processing . . . 44
4.4.3 Common Implementation Details for all Experiments . . . 44
4.4.4 List of Experiments and Specific Implementation Details . . . 45
4.4.5 Results . . . 46
4.5 Discussion . . . 47

5 A Test-Time Adaptation Approach for Robust Medical Image Segmentation 50
5.1 Introduction . . . 50
5.2 Background . . . 52
5.2.1 Denoising Autoencoders . . . 52
5.3 Method . . . 53
5.3.1 Splitting Parameters into Image-Agnostic and Image-Specific . . . 53
5.3.2 Supervised Learning on Training Distribution . . . 54
5.3.3 How to Drive Adaptation at Test-Time? . . . 55
5.3.4 DAE Training . . . 58
5.3.5 Atlas initialization for TTA for large DS . . . 59
5.3.6 Integrating 2D Segmentation CNN with 3D DAE . . . 60
5.4 Experiments and Results . . . 61
5.4.1 Datasets . . . 61
5.4.2 Pre-processing . . . 62
5.4.3 Common Implementation Details for all Experiments . . . 63
5.4.4 List of Experiments and Specific Implementation Details . . . 64
5.4.5 Results . . . 68
5.5 Discussion . . . 76

6 A Task-Agnostic Test-Time Adaptation Approach for Robust Medical Image Analysis 80
6.1 Introduction . . . 80
6.2 Background . . . 83
6.2.1 Markov Random Fields . . . 83
6.2.2 Field of Experts . . . 84
6.3 Method . . . 84
6.3.1 Splitting Parameters into Image-Agnostic and Image-Specific . . . 84
6.3.2 Supervised Learning on Training Distribution . . . 85
6.3.3 FoE-CNN: A New Helper Model for TTA . . . 86
6.3.4 How to Drive TTA using the FoE-CNN Model? . . . 88
6.3.5 FoE-CNN-PCA: An Extended Helper Model . . . 90

6.3.6 How to Drive TTA using the FoE-CNN-PCA Model? 90
6.4 Image Segmentation Experiments and Results 91
6.4.1 Datasets 91
6.4.2 Pre-Processing 91
6.4.3 Common Implementation Details for all Experiments 92
6.4.4 List of Experiments and Specific Implementation Details 93
6.4.5 Results 96
6.5 Image Registration Experiments and Results 101
6.5.1 Datasets 102
6.5.2 Common Implementation Details for all Experiments 102
6.5.3 List of Experiments and Specific Implementation Details 102
6.5.4 Results 103
6.6 Discussion 103
6.6.1 Strengths of TTA-FoE 105
6.6.2 Limitations of TTA-FoE 105
6.6.3 Avenues for further exploration 106
6.7 Appendix 108
6.7.1 Matching Full FoE Distribution v/s Matching Individual Expert Distributions 108
6.7.2 Incorporating Information from Different Training Images 109
6.7.3 TTA-AE variants 109

7 Discussion **111**
7.0.1 Which parameter subset to adapt for the test distribution? 111
7.0.2 DS due to population-based selection bias 113
7.0.3 DS due to task-specific selection biases 114
7.0.4 Acquisition-related DS in image enhancement problems 116
7.0.5 DS due to imaging artifacts or presence of disease 117
7.0.6 DS robustness of density estimation models 117

List of Figures

1.1 Examples of DS in the input space due to acquisition differences 9
1.2 Example performance degradation due to acquisition-related DS 10
1.3 Examples of DS in the input space due to bias with respect to anatomical variations 11
1.4 Examples of DS in the input space due to bias with respect to corruption patterns 11

3.1 Image segmentation datasets - prostate 36
3.2 Image segmentation datasets - heart 36
3.3 Image segmentation datasets - spine 37
3.4 Image segmentation datasets - brain (healthy anatomy) 37
3.5 Image segmentation datasets - brain (white matter hyperintensities) 37
3.6 Image registration datasets 38

4.1 Representative CNN for image segmentation 41
4.2 TL-BN - Qualitative results 48

5.1 TTA-DAE - Proposed parameter split 55
5.2 TTA-DAE - Workflow 56
5.3 TTA-DAE - Qualitative results 70
5.4 TTA-DAE - Comparison with DAE-based post-processing 75
5.5 TTA-DAE - Evolution of predicted segmentation during TTA iterations 75
5.6 TTA-DAE - Convergence behaviour 76
5.7 TTA-DAE - Correlation between supervised and TTA loss 76

6.1 TTA-FoE - Illustration of DS Problem in TTA Helper Models 81
6.2 TTA-FoE - Illustration of Task Specific Experts of TTA-FoE-CNN Model 85
6.3 TTA-FoE - Quantitative results 98
6.4 TTA-FoE - Evolution of predicted segmentation during TTA iterations 99
6.5 TTA-FoE - KDE v/s Gaussian approximation of task-specific expert distributions 100
6.6 TTA-FoE - Workflow for image registration experiments 101
6.7 TTA-FoE - Image registration datasets 102
6.8 TTA-FoE - Subject-wise registration results 104

List of Tables

1.1 Common tasks in medical image analysis . . . 3
1.2 Distribution shifts relevant in medical imaging . . . 7

2.1 Mathematical notation . . . 14
2.2 Machine learning settings for tackling distribution shifts . . . 16

3.1 Datasets - Regions to be segmented for different anatomies . . . 30
3.2 Datasets - Acquisition institutions . . . 31
3.3 Datasets - Scanner and resolution differences . . . 32
3.4 Datasets (Spine) - Acquisition protocol details . . . 34
3.5 Datasets (WMH) - Acquisition protocol details . . . 35

4.1 TL-BN - Dataset details . . . 44
4.2 TL-BN - Quantitative results . . . 46
4.3 TL-BN - Analysis experiments . . . 48

5.1 TTA-DAE - Dataset details . . . 62
5.2 TTA-DAE - Quantitative results . . . 71
5.3 TTA-DAE - Comparison with UDA . . . 72
5.4 TTA-DAE - Ablation study . . . 73

6.1 TTA-FoE - Dataset details (segmentation) . . . 92
6.2 TTA-FoE - Quantitative results (segmentation) . . . 97
6.3 TTA-FoE - KDE vs Gaussian approximation of FoE experts . . . 100
6.4 TTA-FoE - Effect of weighting parameter for the CNN and PCA experts . . . 101
6.5 TTA-FoE - Quantitative results (registration) . . . 104
6.6 TTA-AE - Hyper-parameter tuning . . . 110

Chapter 1

Introduction

1.1 Motivation for Algorithmic Analysis of Medical Images

A lot of medical images, not enough radiologists

Imaging plays an important role in several medical diagnoses, screenings as well as interventions [1]↑. Indeed, from a technological standpoint, medical imaging is one of the most important components of modern healthcare systems. This is reflected in the fast-increasing number of imaging exams conducted in high-income countries. For instance, the number of Magnetic Resonance Imaging (MRI) exams in United States of America more than doubled between the years 2000 and 2015 [2]↑. While such images possess a wealth of information about the patient's health, interpretation of the images is a complex task [3]↑; experts undergo several years of specialized training [4]↑. Even so, high variability exists in image interpretation by multiple experts [5]↑ [6]↑, as well as by the same expert at different time-points [7]↑. Thus, potential benefits of algorithmic image interpretation include (1) reducing the time spent by an expert per image, (2) shortening the training duration of expert radiologists, (3) reducing interpretation errors due to fatigue and (4) quantifying interpretation uncertainty based on acquired signals.

Not enough images, not enough radiologists

On one hand, the number of image acquisitions balloons in high-income countries. On the other hand, due to economic inequity, more than half of the planet's population lacks access to even basic imaging modalities such as x-rays and ultrasound [8]↑. Development of affordable imaging machines [9]↑, aided with automated interpretation algorithms and telemedicine [10]↑, has the potential to improve global health immensely.

Humans are not good at some image analysis tasks

So far, we described how automation in image interpretation can assist human experts to reduce their workload and fasten the interpretation process. For some image tasks such as aligning images (image registration), or assembling pixel representation of images from acquired signals (image reconstruction), algorithms far exceed human performance. Automation may also enable acquisition of low-quality images (with benefits such as reduced acquisition time or harmful dosage given to the patient), followed by image enhancement to obtain the same information as the corresponding high-quality image. Further, if automated image interpretation is fast enough, it can provide feedback for improving image acquisition in the patient-specific manner, thus opening up the possibility of iterative acquisition and interpretation.

Integration of data from different sources

Human experts typically combine information from multiple sources before arriving at a clinical decision. These sources may include images from different modalities as well as non-imaging sensors. Within the realm of image analysis, algorithms may be more adept that humans to systematically integrate information from multiple imaging modalities.

What attributes should an useful image analysis tool have?

To meet the described spectrum of clinical needs, an useful image analysis tool should ideally have the following attributes.

- **Accuracy**: It must either meet human performance at the task at hand, exceed it or entail a human-in-the-loop approach with clear time savings over manual analysis. Error tolerance in medical applications is typically quite low; the tool must pass strict regulatory hurdles before it is deemed suitable for clinical use.

- **Reliability**: It must be able to perform the required analysis under varying quality of the acquired images, and potentially also in the presence of imaging artifacts. When it fails, it must do so in a predictable manner, and provide an indication of failure that would enable a human expert to override the analysis.

- **Speed**: In time-sensitive applications (e.g. surgery), the analysis must be done in a timely manner. On the other hand, speed may not necessarily be the most pressing concern in applications where there exists a delay between the image acquisition and the associated decision-making for other reasons.

1.1.1 Common Tasks in Medical Image Analysis

Recognizing the value that automated image analysis can bring to the clinics, the research field of medical image analysis was born in the late 1970s. [11]↑, [12]↑provide excellent overviews of the initial progress in the field. Below, we enlist clinical tasks have been widely identified as those that can benefit from automation.

- **Image segmentation** [13]↑ [14]↑ [15]↑is the task of delineating anatomical organs or other regions of interest (e.g. lesions). Often, the goal of image segmentation is to measure the organ volume (e.g. to track growth over time) or to locate it precisely (e.g. for therapy planning).
- **Image registration** [16]↑ [17]↑ [18]↑is the task of aligning two images. Typically, such alignment is sought between one of the following pairs: (1) two images of the same patient, acquired at different time-points, (2) two images of the same patient, acquired using different imaging modalities (each potentially highlighting different information about the patient's health), (3) an image of a patient and an atlas.
- **Image classification** [19]↑is the task of categorizing images into a discrete number of classes; for instance, predicting whether an image consists of diseased or healthy anatomy, or whether a tumour is benign or malignant.
- **Image reconstruction** [20]↑, [21]↑is the task of transforming acquired signals (depending on the imaging modality) into interpretable spatial images, potentially in the absence of exact analytical solutions.

Task	Input	Output
Segmentation	Image	Per-pixel labels
Registration	Pair of images	Deformation field
Classification	Image	Category
Reconstruction	Acquired signals	Spatial Image
Super-resolution	Low-resolution image	High-resolution image
Deblurring	Blurred image	Deblurred image
Denoising	Noisy image	Denoised image
Synthesis	Image of one modality	Image of another modality

Table 1.1: Common tasks in medical image analysis.

- **Image enhancement** refers to the task of transforming low-quality images into high-quality ones - for instance, by improving the image resolution [22]↑, sharpness [23]↑or by removing artifacts [24]↑.
- **Image synthesis** [25]↑is the task of transforming images of one modality into corresponding images of another modality; potentially, such a transformation may help to reduce adverse side-effects of the acquisition process of a particular modality.

1.2 Methods for Medical Image Analysis

A common framework for posing all the tasks described in Sec. 1.1.1 is that of input-output transformations. The input and output entities corresponding to different tasks are summarized in Table 1.1. The goal of image analysis algorithms is to construct models that faithfully describe these input-output relationships. Such models can be broadly categorized into two types: hand-crafted and data-driven [26]↑, [27]↑.

1.2.1 Hand-crafted Models

Hand-crafted models leverage a-priori human understanding about the input-output relationship to mathematically define the desired transformation up to a small number of unknown parameters. Here, we describe a common framework, showcasing the application of such models. In this framework, a so-called energy function that depends on the input image, the predicted output and the model parameters, is minimized. The energy consists of two terms:

1. Compatibility of the input and the predicted output. This term is often referred to as data consistency, and its formulation is dictated by our understanding of the input-output relationship. Some examples: for MR image reconstruction from undersampled measurements, the data consistency term leverages knowledge of the noise distribution in the acquired k-space data [28]↑; for image registration, it seeks to increase the similarity between the aligned images [29]↑; for image segmentation, it imposes realistic assumptions on the distribution of image intensities for pixels of each organ [30]↑.
2. Plausibility either of the function mapping the input to the output, or of the predicted output itself. This term is often referred to as regularization, and its formulation is dictated by our understanding of plausible functions /

outputs for the task at hand. Some examples: reconstructed images may be sought to be sparse in a suitable representation [31]↑; deformation fields defining image registration may be required to be diffeomorphic [32]↑; predicted segmentation maps may be encouraged to be spatially smooth [33]↑.

A large number of such methods have been developed, with encouraging results for several medical image analysis tasks. Thanks to the aforementioned parameter optimization is done for each image, such methods tend to work robustly across clinically pertinent changes in imaging protocol parameters. On the other hand, the per-image optimization also makes them slow. Further, the individual terms of the energy function have to be designed for each task, and yet, such methods have managed to provide clinically desirable accuracy only for a small number of tasks.

1.2.2 Data-Driven Models

In many tasks of interest, the input-output relationships of interest entail complex physical processes, such that mathematical specification of model families that capture such relationships is difficult. In such scenarios, data-driven models seek to extract statistical correlations between inputs and outputs from example input-output pairs, which may be relatively easier to generate.

Supervised Learning via Empirical Risk Minimization

A common framework here is to specify the output as a parametric transformation of the input. The parameters of the transformation are then obtained by minimizing a suitable measure of discrepancy between the predicted and true outputs, thus leveraging the aforementioned input-output example pairs. This process is known as empirical risk minimization (ERM) [34]↑. The models are said to have been learned or trained from data. The dataset used for training the models is referred to as the training dataset.

Two characteristics about data-driven supervised learning models are particularly noteworthy: 1. In this setup, the information about the input-output correlations are condensed in the learned model after training. At test time, the trained model parameters are typically fixed, and directly used for making predictions (unlike the per-test-image optimization described in Sec. 1.2.1). 2. Theoretical results from statistical learning theory [35]↑provide performance guarantees for the trained models, in terms of the complexity of the chosen parametric model family and the number of data examples available for training.[1] However, such guarantees hold only as long as the training and test

---

[1] Such results are usually stated as probably approximately correct (PAC) conditions: the

images are sampled from the same probability distribution. In other words, unless additional techniques are employed, the trained models cannot be expected to generalize to out-of-distribution test images. The implications of this behaviour are further expounded in Sec. 1.3.

Deep Learning

A large number of parametric model families have been developed for data-driven supervised learning. Deep learning models are a particular category among the proposed model families. Here, the parametric transformation is modelled via deep neural networks - stacked linear transformations, interlaced with element-wise non-linearities. For data with grid-like data like images, the use of convolutions for parameter sharing within the linear transformations has been shown to be highly effective [36]↑, [37]↑. Further, innovations in the design of the interlacing non-linear functions [38]↑, stochastic optimization algorithms [39]↑and the introduction of specific normalization subroutines [40]↑have facilitated efficient training of deep convolution neural networks (CNNs). Finally, improved computational capabilities [41]↑and availability of data have also played a crucial role in the success of deep learning approaches.

Due to these factors, CNN-based methods are the state-of-the-art in several challenges [42]↑[43]↑[44]↑, often outperforming more traditional methods by large margins in accuracy and applicability to multiple problems [45]↑[46]↑[47]↑. Furthermore, as discussed in Sec. 1.1, human experts often exhibit substantial variability in their interpretation of images [6]↑. For some tasks, anatomies and imaging modalities, the performance of CNN-based methods is already comparable to inter-expert variability [48]↑.

Spurred on by the promising performance of CNN-based analysis methods in a large number of research studies, there have been initial attempts to integrate them within the clinical workflow [49]↑[50]↑[51]↑↑.

---

probability that the prediction error is greater than a certain threshold is bounded from above by a quantity that depends on the error threshold and a measure of complexity of the parametric model family. Recent successes in the setting when the model family is a deep neural network are characterized by very low prediction errors, often even surpassing those guaranteed by theoretical bounds.

1.3 Robustness to Distribution Shifts

1.3.1 The Distribution Shift Problem

An important prerequisite for the performance guarantees of all data-driven models learned using supervised learning (and thus, for CNN models as well) is that the probability distributions of training and test data should match (Sec.1.2.2). When this condition is not met, the predictions of the learned models are unreliable and may demonstrate substantial performance degradation. In other words, while CNNs excel in expressing input-output mappings within the probability distribution corresponding to the training set, they are notorious for responding unpredictably to out-of-distribution inputs - that is, test images that are derived from a different probability distribution [52]↑. This is known as the distribution shift (DS) problem. (We use the acronym DS to refer to both, the singular 'distribution shift' and the plural 'distribution shifts', and call on the reader to infer the form based on the context.) DS occur due to a variety of reasons (Sec. 1.3.2), and are pervasive in clinical practice. Therefore, tackling them suitably is crucial for large-scale adoption of deep learning methods.

1.3.2 Distribution Shifts in Medical Imaging

Several types of DS are pertinent in medical imaging [53]↑. Depending on the factors of the joint probability distribution of the inputs and outputs that change or remain the same across training and test time, the relevant DS can be grouped as described below and summarized in Table 1.2.

Shifted distribution	P(X)	P(Y)	P(X|Y)	P(Y|X)
Causes	Acquisition process Selection bias (anatomical) Selection bias (corruptions)	Label prevalence	Label manifestation	Annotation difference

Table 1.2: Distribution shifts relevant in medical imaging. X and Y indicate inputs and outputs, respectively.

Shift in the conditional distribution of outputs given inputs

For tasks such as segmentation and classification, training datasets consist of input-output pairs, where the task-specific outputs are annotated by domain experts (e.g. radiologists). Due to inherent ambiguity in the task, the annotations made by different experts may vary [6]↑. If the performance of a learned model trained using annotations of one expert is evaluated based on another expert's annotations, the model may exhibit degraded performance.

Shift in the conditional distribution of inputs given outputs

Consider a classification task where an image must be classified as healthy or diseased. If the disease in question manifests in a different way in the test dataset, as compared to the training images, the trained model cannot be expected to make correct predictions at test time.

Shift in the output distribution

Again, consider a classification task where an image must be classified as healthy or diseased. If the proportion of healthy to diseased images in the training and test datasets differ substantially, the biases learned in the trained model may be potentially unsuitable for providing high prediction accuracy at test time.

Shift in the input distribution

In tasks such as segmentation, registration and classification, the inputs to the prediction model are images or pairs of images and outputs are task-specific (Table 1.1). Other tasks such as reconstruction, super-resolution, deblurring, denoising can be considered as image enhancement tasks. Here, the inputs are task-specific corrupted images and outputs are the corresponding corruption-free images. In the former set of tasks, the input images are influenced by two main factors - (1) the acquisition process and (2) the subject's anatomy. In the latter set, the input images are additionally also influenced by (3) the type of corruptions (e.g. undersampling artifacts, type of resolution sub-sampling or blurring kernel, noise distribution) observed during training. Accordingly, the causes of shifts in the input distribution can be categorized as follows. Shifts in the input distribution are also known as covariate shifts in the machine learning literature.

1. **Image acquisition process**: The image acquisition process is the metaphorical lens through which we observe the underlying subject anatomy. Different imaging modalities acquire information about anatomical structure by exploiting different physical processes. This affects the contrasts between two anatomical structures of the same individual, seen in images of different modalities. Indeed, it is possible that a certain imaging modality may starkly differentiate between two structures, which may appear identical to another modality. Clearly, changing the imaging modality (e.g. from CT to MRI) between training and test images represents a distinct DS.

   Further, within the same modality, contrast between different tissues can

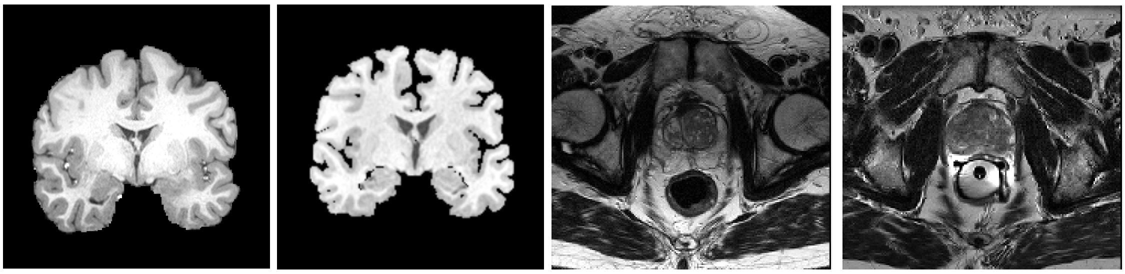


Figure 1.1: Example images showing acquisition-related DS in the input space. The two images on the left show are T1-weighted brain MRI slices from different institutions, while the two images on the right are T2-weighted prostate MRI slices from different institutions. Note that such DS manifest primarily as contrast differences.

be substantially modified by varying the parameters of the acquisition protocol. For instance, in MRI, contrast variations may arise when protocol parameters such as the field strength, resolution, flip angle, echo or repetition time, etc. are changed. In practice, such variations in protocol parameters may be necessitated to obtain the optimal information; as decided in a case-specific manner by the clinical expert conducting the imaging exam.

Even for standardized acquisition protocols, differences between training and test images might still arise due to the usage of different scanners in different acquisition centers. This can happen due to factors such as the drift in scanner SNR over time [54]↑, gradient non-linearities [55]↑, among others. This is further substantiated by the detection of site-specific signals in images even after intensity harmonization [56]↑.

Acquisition-related input DS manifest largely in the form of differences in low-level intensity statistics and contrast changes between different tissue types. Figure 1.1 shows example images of such DS, for different anatomical regions. Evidently, CNNs trained for segmentation rely on such low-level intensity characteristics, thereby demonstrating remarkably degraded performance when confronted with such variations at test time [57]↑. Such lack of robustness of CNN-based methods to such input DS is well-documented in the literature for a number of medical image analysis tasks, as well as in different imaging modalities: lesion segmentation from brain MRIs [58]↑, brain healthy tissue segmentation in MRIs [59]↑, cardiac organ segmentation in MRIs [60]↑, sex classification in MRIs [56]↑, pneumonia classification from chest x-ray images [61]↑[62]↑[63]↑. An example of such performance degradation is shown in Fig. 1.2.

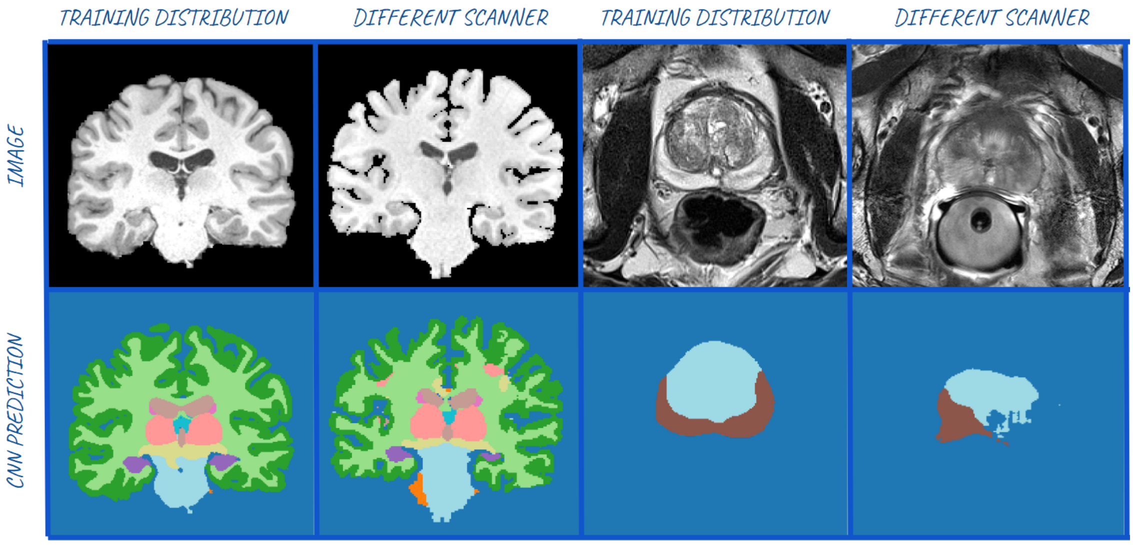


Figure 1.2: Example performance degradation due to acquisition-related DS. A trained CNN provides accurate segmentations for test images from the training distribution, but inaccurate segmentations for test images from a different institution.

2. **Data selection bias in anatomical variations**: This refers to the scenario where the training dataset is not representative of all anatomical variations that may be encountered at test time. Some examples: the training dataset is biased toward a particular demographic (e.g. age, gender [64]↑); the training dataset contains healthy images but images of diseased patients are encountered at test time; the training dataset contains artifact-free images but images with acquisition artifacts are encountered at test time. Examples of such DS are shown in Fig. 1.3.

3. **Data selection bias in corruption patterns**: This scenario is particularly relevant for image enhancement tasks. Here, the training dataset is often retrospectively assembled by corrupting a set of enhanced images using particular corruption patterns (e.g. undersampling patterns, sub-sampling or deblurring kernels, additive noise, etc.). Examples of such differences are shown in Fig. 1.4. Test images corresponding to different corruption patterns than those simulated during training represent an input DS that may hamper prediction performance.

1.4 Contributions

The goal of this thesis is to improve robustness of deep learning-based supervised learning methods for medical image analysis, with respect to acquisition-

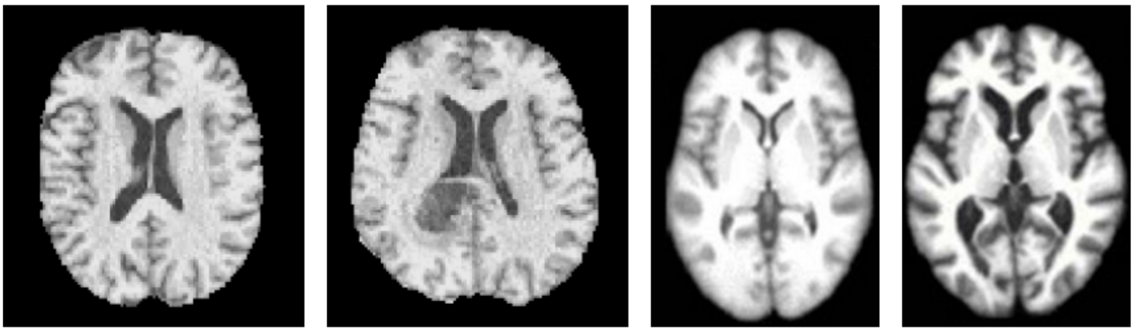


Figure 1.3: Example images DS in the input space due to bias with respect to anatomical variations. The first two images show a selection bias from healthy to diseased - that is, the training dataset consists of images (such as the first one) from only healthy individuals, while an image with a tumour (such as the second one) is encountered at test time. The last two images (taken from [65]↑) show a selection bias with respect to age. The third image is a template of ages 25-29, while the fourth image is a template for ages 85-89.

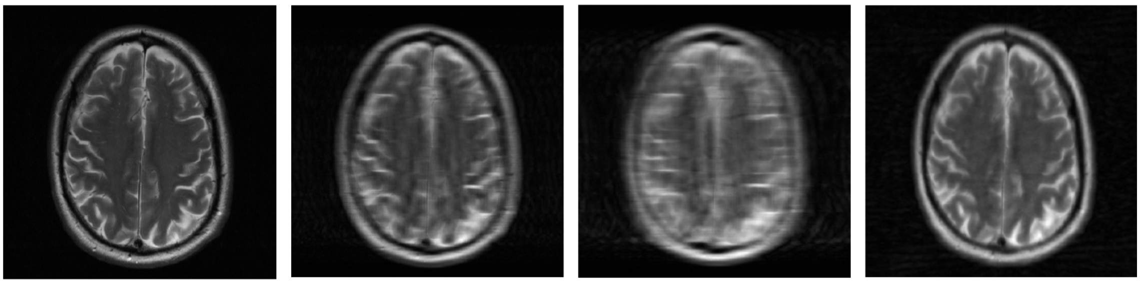


Figure 1.4: Example images for the MRI reconstruction problem, showing DS in the input space due to bias with respect to corruption patterns. From left to right: a fully sampled MR image of the brain, followed by zero-filled reconstructions of undersampled images with cartesian undersampling with factor 4, cartesian undersampling with factor 8 and radial undersampling with factor 8.

related distribution shifts in their inputs. To this end, we have developed the following methods.

[Contribution 1] A transfer learning approach for MRI segmentation

First, we developed a transfer learning method for segmentation of brain MRIs from different scanners and protocols [66]↑. The main idea of this work was to adapt a small number of parameters using a few labelled images at each new distribution.

Experiments showed that the proposed method improved performance on the test distribution substantially, providing comparable results to an independent CNN trained exclusively on a large dataset in the test distribution. Further, the method also had desirable properties from a lifelong learning perspective - performance on the original training distributions was preserved even after model adaptation for the test distribution.

[Contribution 2] A test-time adaptation approach for MRI segmentation

Second, we considered the more practical setting where the test distribution (e.g. hospital) is unknown at training time. As well, transfer of labelled data from the training to the test site was disallowed in view of privacy or technical concerns. In this challenging setting, domain generalization (DG) is the common approach in the field. To further improve the performance of DG methods, we proposed one of the first test-time adaptation (TTA) works [67]↑.

In this work, we deviated from the standard training-testing binary, wherein models are trained during a training phase, fixed thereafter, and directly used to process test images. Instead, we argued, that in order to achieve robustness to unseen variations at test-time, we must allow for model adaptation for each test image. We drove such adaptation by requiring that the segmentations predicted by the model be anatomically plausible. In other words, we utilized an implicit prior in the output space (modelled by a denoising autoencoder). We formulated the segmentation model as a concatenation of a shallow normalization module that was adapted for each test image, and a deep convolutional neural network that segmented the normalized image. We designed the normalization module such that adapting it allowed for contrast variations without substantial structural changes.

Experiments on three anatomies (brain, prostate and heart) demonstrated the viability of TTA as a generic tool to improve segmentation performance for completely unseen test distributions. For some test distributions, thanks to the proposed per-test-image adaptation, TTA lead to even higher performance than that of a CNN trained using supervised learning specifically for that distribution. Further, analysis experiments showed that the design choice of restricting the adaptable module to a shallow normalization sub-network was crucial for good TTA performance.

[Contribution 3] A task-agnostic test-time adaptation approach

Third, we generalized the TTA approach described such that it could be used for tackling DS for multiple tasks. This was done by driving the adaptation with a prior in the feature space, rather than one in the output space. Further, we noted that the CNN-based priors suggested in [68]↑[67]↑may themselves suffer from unreliable behaviour if faced with DS in their inputs. To overcome this problem, we considered a field of experts prior formulation, where the individual experts were modelled to be 1D marginal distributions of the CNN features.

For image segmentation, extensive experimentation on datasets from 17 institutions for 5 tasks (prostate, heart, spine, healthy brain structures, white matter hyperintensities in brain images) revealed that several recent TTA methods provided comparable performance. However, the performance of the FoE-based TTA was the most stable, indicating the FoE model's better DS robustness as compared to helper models used in other works. Further, the method could also improve DS robustness for the task of image registration, while several other methods in the literature could not be applied.

To summarize, we have developed methods with increasing generality to tackle acquisition-related DS in supervised learning CNN models for medical image analysis.

1.5 Layout of the Thesis

The layout of the rest of the thesis is as follows. We present a review of related works in chapter 2. Following this, we describe the datasets used in our experiments in chapter 3. This is followed by a description of our three contributions in chapters 4, 5 and 6. Chapter 7 provides a discussion of our contributions, their relationship with relevant methods in the literature and an outlook for further work.

Chapter 2

Literature review

2.1 Notation

Let X and Y be random variables denoting inputs and outputs, respectively. Let us assume access to a labelled training dataset of paired inputs and outputs, $\mathcal{D}_{tr}^{L}$: $\{(x_i, y_i)| \;\; i = 1, 2, \dots N_{tr}\}$. Here, $x_i \sim P_{tr}(X)$ are samples from the training distribution of inputs and y_i are corresponding ground truth outputs.

Let us consider methods where a deep convolutional neural network (CNN), T_Θ, is used to learn the mapping from X to Y, with Θ indicating the learnable parameters of the CNN. The learning leverages $\mathcal{D}_{tr}^{L}$ by employing a loss function $\mathcal{L}_{task}$ in the empirical risk minimization framework (Sec. 1.2.2). Further, other loss functions may be used to improve robustness with respect to distribution shifts. Let us use a generic notation, $\mathcal{L}_{reg}$, to indicate such losses used

Notation	Description
X	Input
Y	Output
$P_{tr}(X)$	Distribution of training inputs
$P_{ts}(X)$	Distribution of test inputs
$\mathcal{D}_{tr}^{L}$	Labelled training dataset
$\mathcal{D}_{ts}^{L}$	Labelled test dataset
$\mathcal{D}_{ts}^{UL}$	Unlabelled test dataset
N_{tr}	Number of training images
N_{ts}	Number of test images
T_Θ	Deep CNN model for the task
$\mathcal{L}_{task}$	Supervised loss for the task
$\mathcal{L}_{reg}$	Generic regularization loss

Table 2.1: Mathematical notation.

for regularization. We defer the introduction of method-specific notation to the point where those methods are described in detail.

After training, the model is asked to make predictions from test inputs sampled from either the training distribution, $P_{tr}(X)$ or a different test distribution, $P_{ts}(X)$. In some settings, a labelled dataset may be available from the test distribution, $\mathcal{D}_{ts}^{L}$: $\{(x_i, y_i)| \;\; i = 1, 2, \ldots N_{ts}\}$. In some other settings, a dataset of only unlabelled inputs may be available from the test distribution, $\mathcal{D}_{ts}^{UL}$: $\{(x_i)| \; i = 1, 2, \ldots N_{ts}\}$. Both $\mathcal{D}_{ts}^{L}$ and $\mathcal{D}_{ts}^{UL}$ consist of input samples from the test distribution, $x_i \sim P_{ts}(X)$.

A summary of the notation is given in Table 2.1.

Note that some terms that are used interchangeably in the distribution shift robustness literature. First, the term 'domains' is often used to indicate distributions. Second, the training distribution and test distribution are often also referred to as source domain and target domain, respectively. Further, in the context of acquisition-related distribution shifts in medical imaging, the training distribution refers to the distribution of images from one or more scanners or acquisition institutions, while a shifted test distribution refers to that of images from another scanner or institution.

2.2 Machine Learning Settings to Tackle Distribution Shifts

Due to its high practical relevance, the DS problem (Sec. 1.3.1) has attracted substantial attention in the research community. In particular, acquisition-related DS have received the most scrutiny.

An axis along which these efforts can be categorized is that of data requirement - a criterion that is of high relevance for the following two reasons. (1) Scarcity of annotated data is a well-recognized problem in medical image analysis. Depending on the task, annotation can be cumbersome and time-consuming. As well, experts who are qualified to provide such annotations are typically highly strained for resources. (2) Sharing of medical data across institutions is a non-trivial task that often requires regulatory and privacy clearances.

In decreasing order of data requirements, methods in the DS literature can be broadly categorized into the following 7 groups: supervised learning, transfer learning, unsupervised domain adaptation, domain generalization, source-free domain adaptation, test-time adaptation and unsupervised learning. Table 2.2 provides a summary of the data requirement and the main algorithmic ideas of these categories, and a detailed literature review of all categories follows. These settings are described in more detail in the following subsections, and methods proposed in those settings are reviewed.

Setup	Training Institution		Test Institution		
	Data	Algorithm	Data	N_{ts}	Algorithm
Supervised Learning	$\mathcal{D}^L_{tr}$	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr})$	$\mathcal{D}^L_{ts}$	Many	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{ts})$
Transfer Learning	$\mathcal{D}^L_{tr}$	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr})$	$\mathcal{D}^L_{ts}$	Few	Init. at Θ^*_{Tr}, $min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{ts})$
Unsupervised Domain Adaptation	-	-	$\mathcal{D}^L_{tr}, \mathcal{D}^{UL}_{ts}$	Many	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr}) \; + \; \mathcal{L}_{reg}(\mathcal{D}^L_{tr}, \mathcal{D}^{UL}_{ts})$
Domain Generalization	$\mathcal{D}^L_{tr}$	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr}) \; + \; \mathcal{L}_{reg}(\mathcal{D}^L_{tr})$	$\mathcal{D}^{UL}_{ts}$	1	-
Source-Free Domain Adaptation	$\mathcal{D}^L_{tr}$	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr}) \; + \; \mathcal{L}_{reg}(\mathcal{D}^L_{tr})$	$\mathcal{D}^{UL}_{ts}$	Many	Init. at Θ^*_{Tr}, $min_\Theta \; \mathcal{L}_{reg}(\mathcal{D}^{UL}_{ts})$
Test-Time Adaptation	$\mathcal{D}^L_{tr}$	$min_\Theta \; \mathcal{L}_{task}(\mathcal{D}^L_{tr}) \; + \; \mathcal{L}_{reg}(\mathcal{D}^L_{tr})$	$\mathcal{D}^{UL}_{ts}$	1	Init. at Θ^*_{Tr}, $min_\Theta \; \mathcal{L}_{reg}(\mathcal{D}^{UL}_{ts})$
Unsupervised Learning	-	-	$\mathcal{D}^{UL}_{ts}$	1	$min_\Theta \; \mathcal{L}_{reg}(\mathcal{D}^{UL}_{ts})$

Table 2.2: Machine learning settings for tackling distribution shifts.

2.2.1 Supervised Learning

From a technical point of view, the most straight-forward solution to the DS problem is to avoid it - that is, to ensure that the training dataset encompasses all input variability likely to be encountered at test time. This can be achieved in one of two ways: (1) by assembling a very large dataset, including annotated images from a large number of acquisition sites or (2) by training an independent model for each new scanner and protocol setting, using a labelled training dataset from that acquisition institution. Thus, in this setting (described in the first row of Table 2.2), the onus of tackling the DS problem is on the strength of the training dataset. Here, a model is trained for the task via supervised learning in the training distribution, and directly used for making predictions in the test distribution.

Difficulty in Creating Large Training Datasets

The task of assembling such large annotated datasets is difficult in medical imaging due to two main reasons. First, as alluded to in Chapter 1, for most tasks of interest, annotation of medical images requires a high level of expertise. And the people who have such expertise are often highly strained for resources. Second, integrating datasets from different institutions necessitates clearance of several privacy and regulatory hurdles. Despite these challenges, initial efforts are underway in this direction, and several relatively large datasets have been created and publicly shared in recent years. These include the UK-Biobank↑, Nightangle Open Science↑, the Fast MRI dataset [69]↑, the BRATS dataset [70]↑ [71]↑, [72]↑, among others. Such large datasets may enable learning robust models of specific tasks for specific anatomies.

Possibility of Shifts beyond Large Datasets

Nevertheless, given the large number of tasks of interest, and the high degree of variability within each task, it is unrealistic that such large datasets can be created for all tasks of interest. Furthermore, even for the tasks for which such datasets are available, it may be possible that further shifts in the input distribution are encountered, beyond the variations that are covered in the training dataset. Such a scenario is plausible due to the large number of imaging parameters, and other causes of variations that are ubiquitous in medical imaging, as well as introduction of new imaging modalities. This motivates the development of methods to either adapt the learned models to new distributions in an efficient way, or to introduce additional training constraints that ensure robust learning. An overview of these methods is provided in the following sub-sections.

2.2.2 Transfer Learning

In the transfer learning (TL) setting (shown in the second row of Table 2.2), a model is learned in a supervised manner using a large labelled dataset in the training distribution, and further adapted using a small labelled dataset in the test distribution of interest. The main intuitive assumption in this setting is that the training and test distributions are closely related - therefore, the corresponding optimal models for the two distributions would be close in the model parameter space. Thus, the model learned in the training distribution is deemed to provide a good starting point for supervised optimization with a few labelled examples in the test distribution. To express this notion of closeness, the adapted model for the test distribution is often said to have been fine-tuned from the model learned on the training distribution.

Efforts in the TL setting have been proposed for several types of models, including support vector machines (SVMs) [73]↑, thus demonstrating the relevance of the DS problem beyond deep learning as well. In recent years, however, deep learning models have provided state-of-the-art results in a large number of tasks, often substantially outperforming other types of models by high margins [42]↑ [43]↑, [44]↑. Accordingly, recent transfer learning literature [74]↑, [66]↑focuses on deep learning models, attempting to extend their in-distribution performance to scenarios with shifted distributions as well.

Which Parameters to Update?

Modern CNNs often consist of a very high number of parameters. If all those parameters are adapted during TL, the adapted model may potentially overfit to the small labelled dataset in the test distribution. Consequently, an important design choice in TL is the answer to the question: "which parameters should be fine-tuned using the small labelled dataset in the test distribution?" To the best of our knowledge, the literature still lacks a principled solution to this question. Indeed, [74]↑report that the optimal subset of parameters to be updated may be dataset dependent. Common choices for this subset are a few initial CNN layers, a few final CNN layers [75]↑or batch normalization parameters [66]↑. Recent work [76]↑proposes test-data-dependent strategies to decide which model parameters should be updated during TL.

The question of creating two subsets of model parameters - sharing one subset across training and test distributions and adapting the other subset for either each new test distribution or each test image - is a recurring theme in this thesis. We provide different answers to this question in different chapters, depending on the considered setting, and provide justification for the same in those chapters.

Intriguingly, it has also been widely reported that TL can also improve test distribution performance, even when the training and test distributions differ substantially. In particular, several works use ImageNet [77]↑pre-trained CNNs as a starting point for learning models for medical image analyses [78]↑. In such works, all the CNN parameters are typically updated, owing to the large shift between training and test distributions.

Links with Self-Supervised Representation Learning

The main goal in the TL setting is to achieve good performance in the test distribution, using a small labelled dataset from that distribution. TL setting leverages learning of a certain task on data from a training distribution to achieve efficient learning on a shifted distribution on the same task. The pre-trained model weights provide as a better initial point for optimization in the test distribution, as compared to random weight initialization.

Another strategy to determine such good starting model parameter configurations is to pre-train the model on test images themselves, but on another pseudo-task. Such pseudo-tasks can be designed such that their training data is readily available from the test images themselves - a strategy known as self-supervised learning. Common pre-training tasks include image reconstruction, image denoising, image inpainting [79]↑, among several others. Yet another initialization strategy that has gained popularity recently is that of contrastive learning [80]↑. Here, positive and negative pairs of images are formed, and the pre-training strategies requires representations of positive image pairs to be similar to one another, while being dissimilar to those of negative images. After pre-training with either of these strategies, the model parameters are updated in a supervised fashion using the small labelled dataset for the task of interest, as in TL.

Links with Continual / Lifelong Learning

Another machine learning setting that is closely related to TL is that of continual or lifelong learning (CL) [81]↑. Here, the objective is to stack learnings of new tasks or of the same task in new distributions on top of one another, such that later learnings can be done more efficiently by leveraging earlier learnings. A common problem observed in lifelong learning in deep learning models is that the training such models on new tasks / in new distributions drastically deteriorates their performance on old tasks / in old distributions - a problem known as catastrophic forgetting [82]↑.

Thus, CL approaches need to meet two requirements simultaneously: (a) efficient learning in new distributions and (b) preservation of previously learned knowledge, while carrying out model adaptation in new distributions. To satisfy both these requirements simultaneously, two types of approaches have been suggested: (1) Data from old training distributions is used jointly, during model adaptation with data from new test distributions for either directly [83]↑or via learned generative models [84]↑. (2) Importance weights for model parameters [85]↑are determined to indicate relevance of the parameters for previously learned distributions. Now, when the model is adapted for a new distribution, parameters that are deemed to be highly important for previous distributions are prevented from substantial change.

2.2.3 Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) completely relieves the requirement of annotating any images in the test distribution. Instead, unlabelled test distribution images are utilized jointly with the labelled training distribution dataset. A rich literature exists in this setting [86]↑, which can be broadly classified into two major categories.

Distribution-Invariant Features

A common setup for solving function approximation tasks is by formulating the overall mapping, $Y = T_\Theta(X)$, as a concatenation of two steps - by first mapping the inputs to an appropriate representation, $Z = N_\phi(X)$, and then mapping the learned representation to the corresponding outputs, $Y = S_\theta(Z)$. In this setup, a common approach followed by several UDA methods is to encourage absence of distribution-specific signals in the learned representation space, Z. A combination of two loss functions, $\mathcal{L}_{task}$ and $\mathcal{L}_{reg}$, is minimized. Here, $\mathcal{L}_{task}$ is a supervised learning loss that depends on the labelled dataset from the training distribution. $\mathcal{L}_{reg}$ is a unsupervised loss that measures the discrepancy in the representations of images from the training and test distributions. Different divergence measures $\mathcal{L}_{reg}$ have been proposed: maximum mean discrepancy [87]↑, f-divergence via an adversarial framework [88]↑, [89]↑.

Image-to-Image Translation

Another approach for UDA is to learn two mappings, similar to those described above, in two independent steps. Such methods typically operate without an intermediate distribution-invariant representation. Instead, a task model is first learned to describe the input-output mapping for the training distribution dataset. Next, when a test distribution is encountered, an image-to-image translation model is learned to map the test images to corresponding training images [90]↑. Predictions for the mapped images can then to be obtained via the pre-trained task model.

In either of the two approaches described above, UDA requires the entire training distribution dataset to be present while carrying out model adaptation for each new test distribution. This can potentially be a severe requirement in medical imaging, where sharing datasets across institutions often requires regulatory and privacy clearances.

Links with Semi-Supervised Learning

Semi-supervised Learning (SSL) is another machine learning setting for reducing reliance on labelled datasets. Here, a small labelled dataset is used together with a large set of unlabelled images from the training distribution to achieve good task performance in the same distribution. Due to the well-known difficulty of annotating large datasets in medical imaging, and the relative ease of obtaining unlabelled images, SSL has attracted considerable interest in the medical image analysis community [91]↑. UDA differs from this setting in the following respects: (1) the labelled training dataset in UDA is typically large and (2) the unlabelled images in UDA are from a shifted distribution. Nevertheless, both settings have the common goal of utilizing unlabelled images in conjunction with a labelled dataset, to achieve good performance for test images akin to those in the unlabelled dataset. Indeed, recent works have pointed out that several strategies that have been proposed in the SSL literature may be applicable in the UDA setting [92]↑.

2.2.4 Domain Generalization

Domain generalization (DG) seeks to learn robust input-output mappings using labelled datasets from one or more training distributions in such a way that the learned mappings are also applicable to images from unseen test distributions. As there is no adaptation step for each new test distribution, the training distribution datasets are not required after the initial training. A trained CNN is transported and used to perform inference without requiring access to a labeled or unlabeled training set. This is advantageous considering the challenges in data sharing in medical imaging - as compared to sharing the training dataset across institutions, it is much easier to transport a trained CNN for usage with images from new test distributions. Thus, from a practical point-of-view, DG is arguably most attractive among all settings for tackling the DS problem discussed so far. Accordingly, this setting has attracted considerable attention in the literature. The proposed works can be broadly categorized into the following three categories.

Distribution-Invariant Features

Similar to UDA, the introduction of a distribution-invariant intermediate representation is also a common DG approach. This approach assumes that a model trained to extract invariant representations from training distributions will also extract similar features from unseen test distributions. Methods vary in the strategies to ensure distribution invariance. For instance, [93]↑introduce a notion of distributional variance to quantify dissimilarity between distributions, and minimize the same. [94]↑achieve distribution invariant features via a separate pre-training step, in which they train a multi-task autoencoder that aims to discover a feature embedding from which all images of multiple training distributions can be reconstructed. [95]↑, [96]↑use f-divergence minimization via an adversarial framework to promote distribution invariance. If the distributions of interest are known to have commonalities in a particular aspect, such cues can be exploited for encouraging distribution-invariance of extracted features. For instance, for extracting invariant representations in the face of acquisition-related DS, ideas from the shape-appearance disentanglement literature [97]↑may be leveraged. Such disentanglement is especially relevant as CNNs have been shown to have a tendency to rely on texture-based representations [98]↑, while acquisition-related DS robustness requires higher reliance of shape-based representations.

A related DG approach is to encourage learning without dependence on distribution specific spurious correlations either by iteratively discarding highly activated features on the training distribution data [99]↑or by employing causal interventions [100]↑.

Data Augmentation

A related approach is to implicitly encourage distribution invariance by expanding the training dataset to include plausible variations that may be encountered at test time. Methods vary in their procedures for generating simulated input-output pairs. These include applying heuristic transformations on the available data [101]↑, exploiting knowledge about the data generation process [102]↑, alternately searching for worst-case transformations under the current task model and updating the task model to perform well on data altered with such transformations [103]↑[104]↑, leveraging multiple training distributions in order to simulate inputs from in-between distributions [105]↑, using random-weighted convolutional filters [106]↑, using random-weighted convolutional networks [100]↑, training with a fully-synthetic dataset of images representing a large degree of morphological, resolution and acquisition parameter variation [107]↑, among many others.

Recent works suggest that the benefits of data augmentation can be explained in a causality [108]↑framework. Specifically, data augmentation can be seen as a tool for simulating interventional data [109]↑, [100]↑.

Meta Learning

Meta-learning based DG approaches [110]↑ [111]↑ [112]↑to simulate the DS problem during the training of the task model. This is done by having meta-train and meta-test distributions during training and requiring that the gradient updates for the meta-train distributions be such that the task loss is also minimized on the meta-test distributions. Like data augmentation, this approach can also be seen to be implicitly encouraging the task to be learned via distribution invariant features.

Invariance / Equivariance to Transformation Groups

The three categories of DG methods described above are all primarily data-driven. A model-based approach for achieving DS robustness is by building invariances or equivariances to specific transformations into the function class describing the input-output relationship. For instance, CNNs are translation equivariant due to the convolution operator. Methods for achieving equivariance to further transformations, such as rotations have been proposed [113]↑, [114]↑, [115]↑, [116]↑. Recently, Transformer-based architectures [117]↑, [118]↑have empirically demonstrated improved DS robustness as compared to

CNNs [119]↑.

Another approach for encouraging invariances (equivariances) to specific transformations is via regularization losses that encourage such behaviour in the local neighbourhood of the training dataset [120]↑, [121]↑. Specifically, such methods encourage the directional derivative of the model to be small along directions that correspond to known distribution shifts.

2.2.5 Test-Time Adaptation

DG methods described above substantially improve DS robustness in CNNs. However, it is acknowledged in the literature that there still remains a gap to the benchmark performance - that is, the performance achieved via supervised learning, using labelled images from the test distribution. Recognizing this limitation, test-time adaptation (TTA) approaches argue for the need of model adaptability to ensure robust performance in the face of unseen distribution shifts. With this motivation, TTA approaches use DG methods to provide a fairly robust trained model, and further improve performance by fine-tuning the model to specifically suit the test image at hand. Importantly, the adaptation at test time is done using solely unlabelled test image, without access to the training dataset. Works in this setting vary along two broad axes - (a) which parameters are adapted at test time and (b) the loss function that is used to drive the adaptation.

Which parameters to adapt at test time?

TTA methods typically modify only a small subset of model parameters at test time. The main motivation for doing so is to retain benefits of the initial supervised learning on labelled training distributions, and to rely on TTA to provide relatively small corrections to account for the DS. Common choices for the subset of model parameters that are adapted at test time include a normalization module in the CNN's initial layers [67]↑, [122]↑, batch normalization parameters throughout the CNN [123]↑and a combination of shallow adaptable modules at different layers in the CNN [68]↑. These choices are either heuristically motivated or hand-crafted according to the DS at hand. For instance, [67]↑argue that acquisition-related DS in medical imaging manifest primarily as contrast changes, and therefore adapt a sub-network capable of modeling such transformations. Recently, [124]↑propose a Bayesian approach for TTA, where all model parameters are adapted at test time, but large deviations from values learned on the training distribution are penalized.

How to adapt model parameters at test time?

In TTA, model parameters are adapted without access either to labels of the test image or to the training distribution dataset(s). Thus, an unsupervised loss has to be defined, that depends only on the unlabelled test image(s) at hand. Such a TTA loss should act as a good proxy for the supervised loss between the test image and its unavailable ground truth label. In particular, as TTA is mainly achieved via gradient-based optimization, the gradients of the TTA loss with respect to the adaptation parameters must well approximate the corresponding gradients with respect to the supervised loss. Proposed choices for the TTA loss include losses of a pre-trained self-supervised networks [67]↑, [68]↑, [122]↑, the entropy of predictions for the test image(s) [123]↑, or task-specific self-supervised losses such as (i) k-space data consistency in MRI reconstruction CNNs [125]↑ [126]↑, (ii) smoothness of predicted segmentations [127]↑, (iii) cycle-consistency-based estimation of a correction filter to transform low-resolution (LR) test images to resemble LR images seen during training of super-resolution CNNs [128]↑or (iv) an estimator (Stein's unbiased risk estimator) of the true loss for known noise distributions in denoising CNNs [129]↑.

Instead of using gradient-based optimization at test-time, [130]↑propose to train via a meta-learning framework, a helper-model that takes test images as inputs and outputs adapted task-model parameters. While previous meta-learning-based DG approaches like [110]↑, [111]↑, [112]↑implicitly encourage learning via distribution-invariant features, this meta-learning based TTA approach [130]↑allows the helper-model to output a different set of parameters particularly suited for the test image at hand. However, the helper model itself is learned during training and fixed thereafter. Therefore, the performance of this approaches relies on similarity between the test distributions seen during training and the test distribution encountered at test time.

Adaptive Batch Normalization

Batch normalization [40]↑(BN) layers have attracted substantial attention in the distribution shift literature, in multiple machine learning settings. [67] adapt BN parameters in a transfer learning setting, using a small number of labelled images from a test distribution. In the TTA setting, [131]↑, [132]↑propose to use the statistics of the test image(s) in the BN layers of the task CNN. Here, no learnable parameters of the task CNN are adapted; rather the mean and variance stored in each batch normalization layer are replaced with those of the given test image(s). Effectively, at each layer, this amounts to matching the 1D Gaussian approximation of the marginal feature distribution of the test image(s) with that of the entire training dataset. Such a strategy has been shown to improve DS robustness in natural imaging datasets. On the other hand, [133]↑point out that this method matches only the first two moments of the 1D distributions, and is thus prone to inaccuracies when the distributions are substantially non-Gaussian.

TTA in Generative Models

Test-image-specific adaptation has also been considered in the context of generative models. For instance, [134]↑propose to fine-tune density estimation models (e.g. generative adversarial networks) for each test image, when used in the Bayesian image enhancement framework. Further, [135]↑report that CNNs trained from scratch to generate a given corrupted test image from a random vector have a tendency to first generate the corresponding clean image. This has been recently leveraged for dynamic cardiac MRI reconstruction in [136]↑.

2.2.6 Source-Free Domain Adaptation

A closely related setting to TTA is that of source-free domain adaptation (SFDA) [137]↑, [138]↑, [139]↑, [140]↑. TTA and SFDA differ in the number of test images used for carrying out the model adaptation in the test distribution. The adaptation is done for each test image in TTA, while multiple images from the test distribution are used simultaneously for model adaptation in SFDA. Apart from this difference, both settings operate by answering the two questions described in the previous sub-section - which parameters to adapt for the test distribution, and how to drive such adaptation by accessing neither labels in the test distribution nor the training dataset. While SFDA has the advantage that multiple images from the test distribution may provide a regularization effect on one another during adaptation, TTA may benefit from adapting parameters to get the best performance for each test image.

2.2.7 Post-Processing Methods

It has been suggested to post-process model predictions, to potentially remove errors due to DS in the inputs. In particular, for image segmentation, such post-processing may be driven by a smoothness prior defined using conditional random fields [141]↑, [142]↑, a prior based on denoising autoencoders [143]↑, or on generative adversarial networks [144]↑. While such post-processing methods may improve plausibility of predictions, they lack a mechanism to ensure that the post-processed prediction is accurate for the given input image.

2.2.8 Unsupervised Learning

The DS problem exists because distributions of training and test images differ. One way to circumvent this problem entirely is to forgo a training step altogether. Without this step, any reliance on distribution-specific signals from a particular training distribution cannot arise. The most successful approaches in this category are based on probabilistic generative models [30]↑, [145]↑, [146]↑. These methods rely on mechanistic models (Sec. 1.2) and pose the problem in a Bayesian framework. For instance, for image segmentation, such methods infer the posterior probability of the unknown segmentation by specifying a prior model of the underlying tissue classes and a likelihood model, potentially, describing the image formation process. A downside of these approaches is that they have so far been largely restricted to prior models encoding similarities in relatively small pixel neighbourhoods [30]↑, [147]↑, [148]↑, [149]↑. Further, they are mainly used in neuroimaging applications where atlas-based approaches are reliable due to limited morphological variation [150]↑, [151]↑. Recent works leverage a set of segmentations in order to learn long-range spatial regularization priors through Markov random fields with high-order clique potentials [152]↑, [153]↑as well as through variational auto-encoders [154]↑. Nevertheless, most of these methods involve deformable image registration as one of their pre-processing steps, thus making it challenging to extend them to applications beyond neuroimaging.

2.3 Problems Related to Distribution Shift Robustness

2.3.1 Out-of-distribution Detection

Out-of-Distribution (OOD) detection refers to flagging inputs derived from a distribution that is shifted from a model's training distribution. This enables the model to acknowledge its inability to correctly process the given test input. Such behaviour is more useful than confidently producing incorrect predictions.

On one hand, major causes of DS in medical imaging are largely known; thus, the existence of such shifts in the input images is likely to be already known when the predictions are made for test images. Nevertheless, OOD detection may still be useful to detect to further shifts in the input distribution - for example, due to imaging artifacts. Such detection can be potentially especially useful in integrated acquisition-analysis systems.

Here, we briefly summarize the main approaches proposed for the OOD detection problem. Please refer to [155]↑for a detailed survey.

One of the main approaches of OOD detection methods is that of density estimation. Predominantly, such approaches learn a model of the training input distribution, and flag test inputs in low density regions under this model as OOD [156]↑. Alternatively, the learned generative model may be asked to generate an image similar to the test image at hand; failure in doing so indicates that the test image is OOD [157]↑. Methods vary in the type of model that is chosen for the training input distribution. For instance, [158]↑fit Gaussian distributions to the task CNN representations, [159]↑use normalizing flow-based [160]↑generative models, [161]↑, [162]↑use energy-based unnormalized probability models, while [163]↑use an ensemble of multiple generative models. [164]↑demonstrate that parametric density estimation models may assign higher likelihood values to OOD samples than samples of the training distribution. To circumvent this issue, [165]↑and [166]↑estimate low dimensional marginal distributions of task-CNN projections non-parametrically, using kernel density estimation.

Other OOD approaches include usage of self-supervised tasks [167]↑, [168]↑, learning OOD classifiers using example OOD inputs [169]↑, [170]↑, temperature scaling of softmax outputs for tasks with categorical outputs [171]↑, among others.

2.3.2 Model Performance Prediction

In the medical imaging literature, an alternative setting to OOD detection is that of model performance prediction (MPP). Here, the goal is to estimate the prediction accuracy of a trained model on test inputs, without access to the corresponding ground truth labels. In other words, OOD detection refers to detecting shifts in the input distribution, while MPP refers to detecting shifts in either the output distribution or in the conditional distribution of the output given an input.

The MPP setting has mostly been explored in the context of image segmentation. One strategy is to train a supervised segmentation prediction regressor based on hand-crafted features from predicted segmentations [172]↑, [173]↑. This idea was extended in [174]↑, where CNN was trained in a supervised manner to predict segmentation accuracy. In a different approach known as reverse classification accuracy prediction, [175]↑use the predicted labels of a test image to train a new segmentation model, the accuracy of which on labelled training examples is used as an estimate of the accuracy of the test image's predicted labels. Another MPP strategy is to obtain uncertainty estimates, with the hypothesis that uncertain predictions are likely to be incorrect [176]↑, [177]↑.

2.3.3 Adversarial Robustness

So far we have discussed degradation in model performance due to naturally occurring shifts in the input distribution between training and test images. Different from this, lack of adversarial robustness refers to the phenomenon that performance degradation can be caused by specific, but imperceptible manual changes in the input to a trained deep learning model. Methods for determining such specific perturbations are known as adversarial attacks, while those seeking to make models robust against such attacks are known as adversarial defences. A large number of such attacks and defences have been proposed in the computer vision literature [178]↑, as well as in the medical imaging literature [179]↑. Further, [180]↑argue that lack of adversarial robustness creates peculiar fraud opportunities that need to be countered via algorithmic as well as infrastructural defenses.

Chapter 3

Datasets

We have developed three methods to tackle acquisition-related distribution shifts in medical image analysis: (i) TL-BN: A transfer learning approach for robust MRI segmentation (chapter 4), (ii) TTA-DAE: A test-time adaptation approach for robust MRI segmentation (chapter 5), and (iii) TTA-FoE: A task-agnostic test-time adaptation approach for robust image analysis (chapter 6). All three methods were validated for the task of image segmentation. Our task-agnostic method, TTA-FoE, was additionally evaluated for the task of image registration as well. We describe the datasets used in our segmentation experiments in Sec. 3.1 and those used in our registration experiments in Sec. 3.2.

Some of these datasets have been publicly available since many years, while some other have been released in recent years. Consequently, different methods developed in the thesis were evaluated on different subsets of the described datasets. The most recently developed method, TTA-FoE, was evaluated on the largest subset, consisting of data from 17 institutions. In that chapter, we also present comparisons the two previously developed methods (TL-BN, TTA-DAE) for all the 17 datasets. However, several analysis and ablation experiments were done for TL-BN or TTA-DAE on smaller number of datasets. This information is provided in the corresponding chapters.

3.1 Datasets for Image segmentation

We considered five segmentation tasks - healthy structure segmentation from four anatomies (prostate, heart, spine, brain), and abnormality segmentation in one anatomy (brain). Table 3.1 lists the regions to be segmented for images of different anatomies.

3.1.1 Reasons for DS in the used Datasets

For each segmentation task, we used images acquired from multiple centers, representing instances of acquisition-related DS. We treat images from each center as samples from separate probability distributions. These distributions may differ due to the following reasons.

I. DS due to variations in population demographics

Table 3.2 lists the names of all the acquisition centers. For each segmentation tasks, the acquisition centers are from different cities, and often from different countries. This could result in potential populated-related distribution shifts (DS), in addition to the acquisition-related DS emanating from scanner and protocol differences.

II. Acquisition-related DS (hardware)

Table 3.3 summarizes the scanner-related differences in the datasets. In the considered datasets, the centers use scanners of one of three vendors: Siemens, Philips and General Electric (GE). Further, even within the same vendor, scanner models and field strengths differ. Both these factors can result in hardware-caused differences in the image statistics.

III. Acquisition-related DS (protocol parameters)

In several cases, there are variations in acquisition protocol parameters (e.g. echo time, repetition time, flip angle, etc.) can be different across centers (e.g. Table 3.4, Table 3.5). These variations further contribute to the DS between centers.

IV. Acquisition-related DS (resolution)

Further, as shown in Table 3.3, the images for the same anatomy are often ac-

Anatomy	# Labels	Foreground labels
Prostate	2	Whole prostate gland
Heart	4	Left ventricle cavity, Right ventricle cavity, Left ventricle myocardium
Spine	3	Spinal cord white matter, Spinal cord grey matter
Brain (Healthy)	15	Cerebellum gray matter, Cerebellum white matter, Cerebral gray matter, Cerebral white matter, Thalamus, Hippocampus, Amygdala, Ventricles, Caudate, Putamen, Pallidum, Ventral diencephalon, Cerebrospinal fluid, Brain stem.
Brain (WMH)	2	White matter hyperintensities

Table 3.1: Regions to be segmented for different anatomies. The number of labels includes all the listed foreground labels, plus one background label for all remaining pixels in the image.

quired with different in-plane and through-plane resolution at different centers. We re-sample the images to make their in-plane resolution consistent across all centers for a particular anatomy. However, the re-sampling may not necessarily resolve the differences in image statistics caused due to acquisition at different resolutions. Thus, the resolution differences may also contribute to the DS between centers.

Figures 3.1, 3.2, 3.3, 3.4 and 3.5 show example images from different datasets for each anatomy. Visually, the DS manifests primarily as a contrast change. Apart from the summaries provided in the Table 3.2 and Table 3.3, anatomy specific additional information for the datasets is given in the following subsections.

Dataset	Center	City	Country	Center ID
	Prostate [181]$^{\uparrow}$[42]$^{\uparrow}$[182]$^{\uparrow}$			
NCI-13	Radboud University Nijmegen Medical Centre	Nijmegen	Netherlands	RUNMC
NCI-13	Boston Medical Center	Boston	USA	BMC
PROMISE12	Haukeland University Hospital	Bergen	Norway	HK
PROMISE12	Beth Israel Deaconess Medical Center	Boston	USA	BIDMC
PROMISE12	University College London	London	England	UCL
PROMISE12	Radboud University Nijmegen Medical Centre	Nijmegen	Netherlands	RUNMC
USZ	Universitaetspital Zuerich	Zurich	Switzerland	USZ
	Heart[183] $^{\uparrow\uparrow}$[43]$^{\uparrow}$[184]$^{\uparrow}$			
M&Ms	Cliınica Sagrada Familia	Barcelona	Spain	CSF
M&Ms	Universitaetsklinikum Hamburg-Eppendorf	Hamburg	Germany	UHE
M&Ms	Hospital Vall d'Hebron	Barcelona	Spain	HVHD
ACDC	University Hospital of Dijon	Dijon	France	ACDC
RVSC	Rouen University Hospital	Rouen	France	RVSC
	Spine [185]$^{\uparrow}$			
SCGM	University College London	London	England	UCL
SCGM	Polytechnique Montreal	Montreal	Canada	PM
SCGM	Universitaetspital Zuerich	Zurich	Switzerland	USZ
SCGM	Vanderbilt University	Nashville	USA	VU
	Brain (Healthy) [186]$^{\uparrow}$[187]$^{\uparrow\uparrow\uparrow\uparrow}$[188]$^{\uparrow\uparrow}$[IXI]$^{\uparrow}$			
HCP	Washington University	St. Louis	USA	HCP
ABIDE	The Adolphs Lab	Pasadena	USA	AC
ABIDE	Stanford University	Stanford	USA	AS
ADNI	Multiple scanners $^{\uparrow}$	Multiple centers $^{\uparrow}$		ADNI
IXI	Multiple scanners $^{\uparrow}$	London	England	IXI
	Brain (White Matter Hyperintensities) [189]$^{\uparrow\uparrow}$			
WMH-17	University Medical Center	Utrecht	Netherlands	UMC
WMH-17	National University Health System	Singapore	Singapore	NUHS
WMH-17	VU University Medical Centre	Amsterdam	Netherlands	VU

Table 3.2: Acquisition institutions of segmentation datasets.

3.1.2 Prostate

Number of Institutions: We used transverse T2-weighted MR images from two publicly available datasets: (i) National Cancer Institute (NCI-13) [181]↑, (ii) PROMISE12 [42]↑and one private dataset: (iii) University hospital in Zurich (USZ) [182]↑. The NCI and PROMISE12 datasets consist of images from dif-

Dataset	Center ID	Vendor	Scanner	Field (T)	In-plane Resolution (mm^2)	Through-plane Resolution (mm)
Prostate [181]↑[42]↑[182]↑						
NCI-13	RUNMC	Siemens	TIM	3	0.5 to 0.75	4.0
NCI-13	BMC	Philips	Achieva	1.5	0.4	3.0
Promise12	HK	Siemens	INA	1.5	0.625	3.6
Promise12	BIDMC	GE	INA	3	0.25	2.2 to 3.0
Promise12	UCL	Siemens	INA	1.5, 3	0.325 to 0.625	3.0 to 3.6
Promise12	RUNMC	Siemens	INA	3	0.5 to 0.75	3.6 to 4.0
Private	USZ	Siemens	Skyra	3	0.5 to 0.75	3.6 to 4.0
Heart [183]↑↑[43]↑[184]↑						
M&Ms	CSF	Philips	Achieva	1.5	1.2	9.9
M&Ms	UHE	Philips	Achieva	1.5	1.45	9.9
M&Ms	HVDH	Siemens	Magnetom Avanto	1.5	1.32	9.2
ACDC	ACDC	Siemens	Trio Tim	1.5, 3	1.34 to 1.68	5.0 to 10.0
RVSC	RVSC	Siemens	Symphony Tim	1.5	0.75 to 1.6	7
Spine [185]↑						
SCGM	UCL	Philips	Achieva	3	0.5	5.0
SCGM	PM	Siemens	TIM Trio	3	0.5	5.0
SCGM	USZ	Siemens	Skyra	3	0.5	2.5
SCGM	VU	Philips	Achieva	3	0.65	5.0
Brain T1w (Healthy structures) [186]↑[187]↑↑↑↑[188]↑↑						
HCP	HCP-T1	Siemens	Skyra*	3	0.7	0.7
ADNI	ADNI-T1	Multiple scanners ↑		1.5, 3	1.0	1.0
ABIDE	AC-T1	Siemens	Magnetom TrioTim	3	1.0	1.0
ABIDE	AS-T1	GE	SIGNA	3	$0.859 * 1.5$	0.859
Brain T2w (Healthy structures) [186]↑[IXI]↑						
HCP	HCP-T2	Siemens	Skyra*	3	0.7	0.7
IXI	IXI-T2	Multiple scanners ↑		1.5	1.75	2.0
Brain (White Matter Hyperintensities) [189]↑↑						
WMH-17	UMC	Philips	Achieva	3	$0.96 * 0.95$	3.0
WMH-17	NUHS	Siemens	TrioTim	3	1.0	3.0
WMH-17	VU	GE	Signa HDxt	3	0.98	1.2
WMH-17	VU	GE	Signa HDxt	1.5	1.21	1.3
WMH-17	VU	Philips	Ingenuity	3	1.04	0.56

Table 3.3: Details of scanner and resolution differences of segmentation datasets. INA stands for 'Information not available'. * The hardware of the standard commercial Skyra scanner was customized [186]↑.

ferent acquisition centers. Images from the center RUNMC are included in both the NCI and PROMISE12 datasets, with NCI-13 providing a higher number of images from this center. Therefore, we ignore this center's data from the PROMISE12 dataset. Thus, we have 6 distributions in the prostate segmentation experiments.

Acquisition-related remarks about particular datasets: The NCI-13-RUNMC and PROMISE12-UCL images were acquired without endo-rectal coils, while the rest of the datasets were acquired with them [190]↑. For the USZ dataset, the acquisition sequence parameters were in accordance with the international prostate MR guidelines [191]↑. For the T2-weighted images considered in our experiments, a turbo spin echo sequence was followed, with echo time (TE) set to 93ms and repetition time (TR) set to 3500ms. For other datasets, the acquisition protocol parameters are not summarized in the corresponding papers, and may differ across images within the same dataset.

Label information: For the NCI-13 and USZ datasets, expert annotations are available for the central gland (CG) and peripheral zone (PZ). For the PROMISE12 dataset, expert annotations are only available for the whole prostate gland (CG + PZ). Fig. 3.1 shows example images from different prostate segmentation datasets.

3.1.3 Heart

Number of Institutions: For cardiac segmentation, we used the publicly available training data from the multi-centre, multi-vendor and multi-disease (M&Ms) cardiac segmentation challenge [183]↑↑. This consists of labelled images from 3 centers. Additionally, we also used the Automated Cardiac Diagnosis Challenge (ACDC) dataset [43]↑and the right ventricle segmentation challenge (RVSC) dataset [184]↑.

Acquisition-related remarks about particular datasets: The images are of healthy individuals as well as subjects with different cardiovascular diseases such as hypertrophic cardiomyopathy, dilated cardiomyopathy, coronary heart disease, abnormal right ventricle, myocarditis and ischemic cardiomyopathy.

Label information: For the M&Ms and ACDC datasets, expert annotations are available for 3 regions: the left and right ventricle (LV and RV, respectively) cavities and the LV myocardium. Further, these annotations are provided at two time-points in the 4D cine images - at the end-diastolic (ED) and end-systolic (ES) phases. We consider each 3D volume as one image - thus, we have two 2 images per subject. For the RVSC dataset, expert annotations are provided for RV cavity and the RV myocardium. Fig. 3.2 shows example images from different heart segmentation datasets.

3.1.4 Spine

Number of Institutions: We used data from the spinal cord grey matter segmentation challenge [185]↑, which provides multi-centre and multi-vendor images acquired with distinct 3D gradient-echo sequences. Images are acquired from 4 centers.

Acquisition-related remarks about particular datasets: The details of the acquisition differences between the datasets are summarized in Table 3.4. Apart from variations in the time parameters, two out of the four datasets were acquired with accelerated MRI techniques. Potentially, the different reconstruction algorithms can act as additional DS sources.

Center ID	Pulse sequence	TE (ms)	TR (ms)	Flip angle (°)	Acceleration
UCL	3D gradient echo	5	23	7	-
PM	2D spoiled gradient multi-echo	5.41, 12.56, 19.16*	539	35	2 (GRAPPA) [192]↑
USZ	3D multi-echo gradient-echo	19	44	11	-
VU	3D multi-echo gradient-echo	7.2, 16.1, 25*	700	28	2 (SENSE) [193]↑

Table 3.4: Acquisition protocol details for the spine datasets. * Averaged offline to create a single image with increased signal-to-noise ratio.

Label information: The regions to be segmented are the grey matter and white matter in the spinal cord. Fig. 3.3 shows example images from different spine segmentation datasets.

3.1.5 Brain (Healthy structure segmentation)

Number of Institutions: For brain segmentation, we used images from 4 publicly available datasets: Human Connectome Project (HCP) [186]↑, Alzheimers Disease Neuroimaging Initiative (ADNI) [188]↑↑, Information eXtraction from Images (IXI) ↑and Autism Brain Imaging Data Exchange (ABIDE) [187]. In the HCP dataset, both T1-weighted (HCP-T1) and T2-weighted (HCP-T2) images are available for each subject. We consider these images are belonging to two different distributions. The ABIDE dataset consists of T1-weighted images from several imaging sites. Of these, we randomly select two sites - California institute of technology (AC-T1) and Stanford university (AS-T1).

Label information: While providing great imaging data in large quantities, unfortunately, these datasets do not provide manual segmentation labels. Moreover, we are not aware of publicly available brain MRI datasets in large quantities with manual segmentations of multiple subcortical structures. In this situation, we employ the widely used FreeSurfer [194]↑tool to generate pseudo ground truth segmentations for the healthy brain structures in all the used datasets. FreeSurfer is a successful segmentation tool, that works robustly across scanner and protocol variations. However, it has the downside of being excessively time expensive, taking as much as 10 hours on a CPU for segmenting one 3D MR image, and specific to the brain. FreeSurfer provides a large number of segmentation labels. We combine these labels into the following 15 regions: background, cerebellum gray matter, cerebellum white matter, cerebral gray matter, cerebral white matter, thalamus, hippocampus, amygdala, ventricles, caudate, putamen, pallidum, ventral diencephalon, cerebrospinal fluid and brain stem. Fig. 3.4 shows example images from different brain healthy tissue segmentation datasets.

3.1.6 Brain (Cerebral White Matter Hyper-intensities)

Number of Institutions: We used data from the White Matter Hyper-intensities (WMH) segmentation challenge[189]↑↑. It consists of data from three centers. T1-weighted and FLAIR (fluid-attenuated inversion recovery) images are provided from each center.

Acquisition-related remarks about particular datasets: In our experiments, we use the FLAIR images only. Images in the UMC and NUHS datasets were acquired with a 2D FLAIR sequence in the transversal orientation, while those in the VU datasets were acquired with 3D FLAIR sequences in sagittal orientation. The details of the acquisition differences between the datasets which are summarized in Table 3.5. In order to restrict the considered distribution shift, we conducted experiments with the two datasets (UMC and NUHS) with 2D sequences only.

Center ID	TE (ms)	TR (ms)	TI (ms)
UMC	125	11000	2800
NUHS	82	9000	2500
VU (3T, GE)	126	8000	2340
VU (3T, Philips)	279	4800	1650
VU (1.5T, GE)	117	6500	1987

Table 3.5: Acquisition protocol details for the WMH datasets.

Label information: White matter hyperintense (WMH) regions of presumed vascular origin [195]↑are to be segmented. Fig. 3.5 shows example images from different brain white matter hyperintensities segmentation datasets.

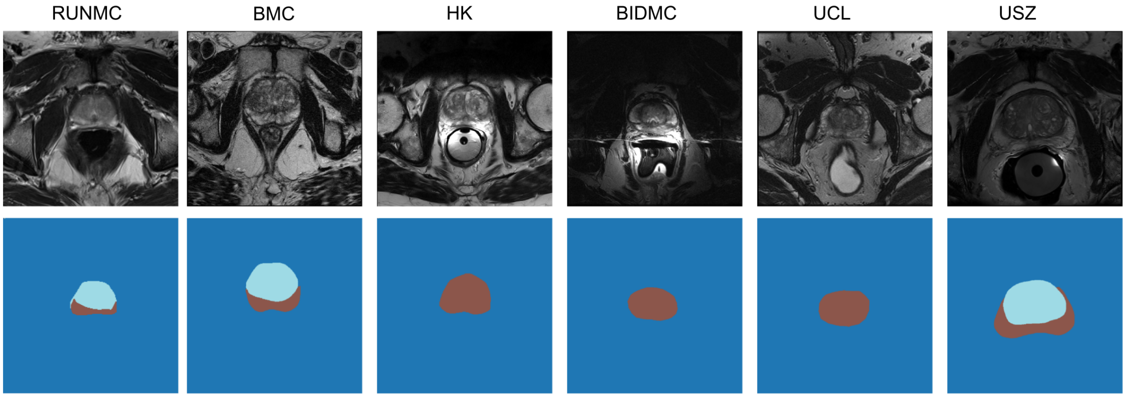


Figure 3.1: Example images and ground truth segmentations from different datasets for prostate segmentation. For the NCI-13 (RUNMC and BMC) and USZ datasets, expert annotations are available for the central gland (light blue) and peripheral zone (brown). For the PROMISE12 dataset (HK, BIDMC and UCL), expert annotations are only available for the whole prostate gland (brown), which includes the central as well as the peripheral zones.

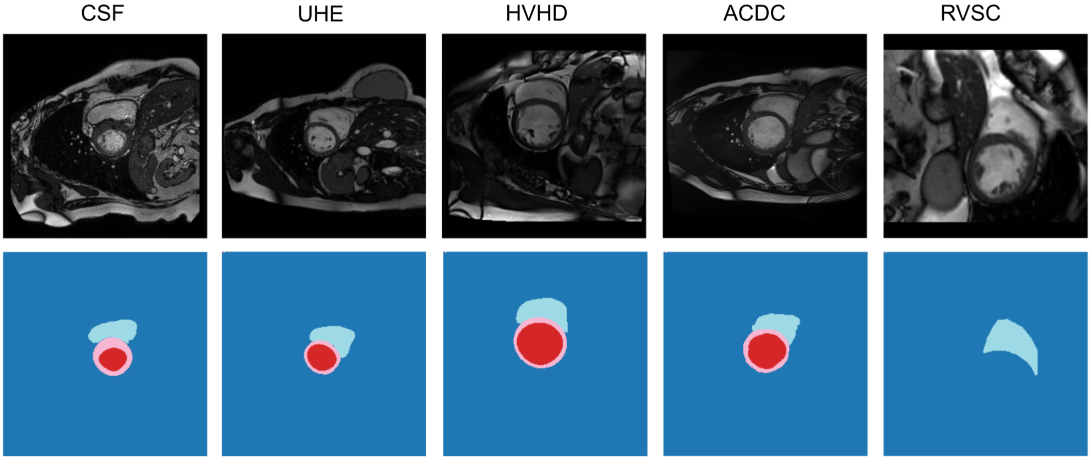


Figure 3.2: Example images and ground truth segmentations from different datasets for heart segmentation. For the ACDC and M&Ms (CSF, UHE, HVHD) datasets, expert annotations are available for 3 regions: the left and right ventricle (LV and RV, respectively) cavities and the LV myocardium. For the RVSC dataset, expert annotations of only the RV cavity is shown, as this is only common region with the other datasets.

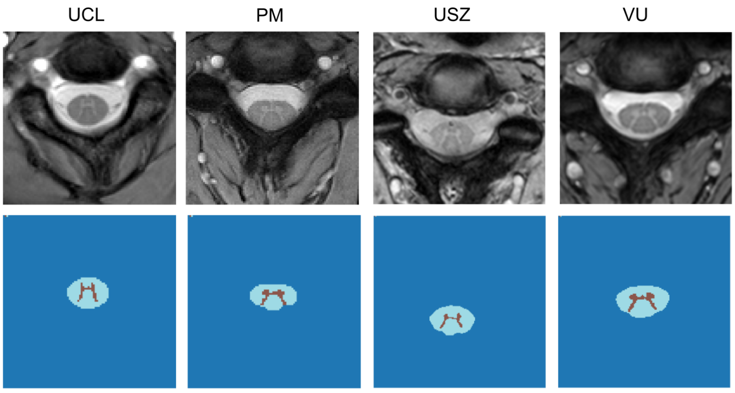


Figure 3.3: Example images and ground truth segmentations from different datasets for spine segmentation.

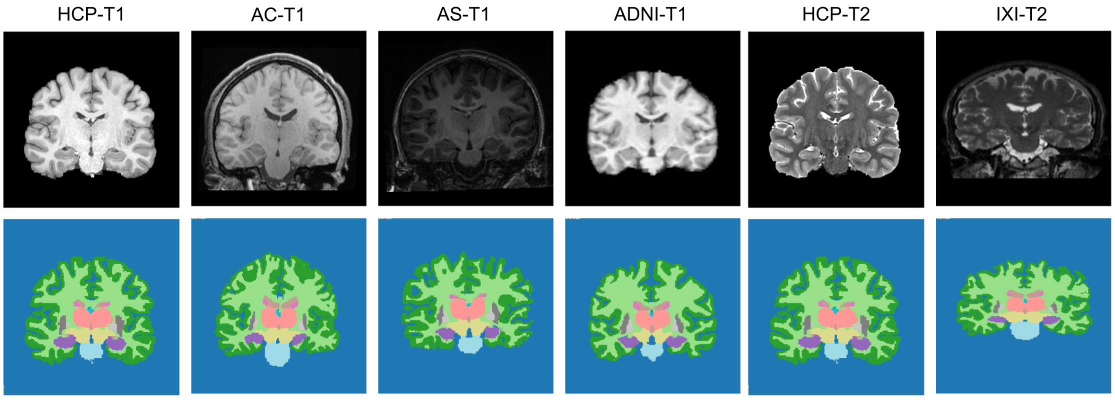


Figure 3.4: Example images and ground truth segmentations from different datasets for brain healthy tissue segmentation.

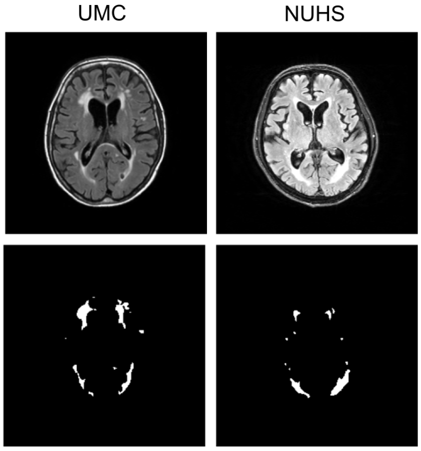


Figure 3.5: Example images and ground truth segmentations from different datasets for brain white matter hyperintensities segmentation.

3.2 Datasets for Image registration

We use images from T1w images from three datasets: HCP [186]↑, ABIDE (AS-T1) [187]↑↑↑↑and OASIS [196]↑. The images are registered with the atlas provided by [197]↑↑. Example slices of the subject images and the atlas are shown in Fig. 3.6.

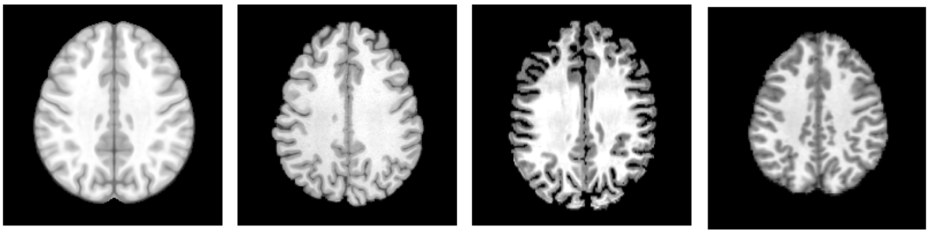


Figure 3.6: From left to right: a 2D slice from the atlas and example slices from three datasets: HCP, ABIDE-STANFORD (AS) and OASIS.

Chapter 4

A Transfer Learning Approach for Robust Medical Image Segmentation

This chapter is based on the publication "A lifelong learning approach to brain MR segmentation across scanners and protocols." [66]↑. Here, we address the lack of robustness of convolutional neural networks (CNNs) to acquisition-related distribution shifts (DS) in a transfer learning setting.

4.1 Introduction

The Transfer Learning Setting

As described in Sec. 2.2, transfer learning refers to fine-tuning parameters of a trained model using a few labelled sampled from the test distribution.

Batch Normalization for Tackling Distribution Shifts

In the computer vision literature, several adaptations of batch normalization (BN) layers [40]↑have been suggested for domain adaptation [198]↑, [199]↑and multi-domain learning [200]↑, [201]↑for object recognition using CNNs. The main idea in these works is to employ BN layers for distribution-specific scaling to account for DS, while sharing the bulk of the CNN parameters to leverage the similarity between the distributions.

Summary of the Proposed Method

In this work, we extend such approaches for segmentation across MRI scanning protocols. Our solution is a single CNN with shared convolutional filters and distribution-specific BN layers, which can be tuned to new distributions with 4 labelled volumetric images (2 for training and 2 for validation). We note that notwithstanding variations in image statistics due to inter-scanner differences, a segmentation network would be confronted with images of the same organ, acquired with the same imaging modality. Thus, it is reasonable to postulate common characteristics between the distributions and consequently, shared support in an appropriate representation space. Following [200]↑, we hypothesize that such a representation space can be found by using distribution-agnostic convolutional filters and that the inter-domain differences can be handled by appropriate normalization via distribution-specific BN modules. On one hand, the proposed approach is in line with previous domain adaptation works [199]↑. On the other hand, it also embodies the normalization idea of conventional proposals for dealing with inter-scanner variations [202]↑, [203]↑, [204]↑. Furthermore, the proposed approach is also attractive from the perspective of lifelong learning [81]↑as well - performance improvement for test distributions is achieved, while retaining performance on the older domains whose training data may no longer be available.

Evaluation
The proposed method is evaluated for brain structure segmentation in MR images. Experiments demonstrate that the method largely closes the gap to the benchmark, which is training a dedicated CNN for each input distribution.

4.2 Background

4.2.1 Batch Normalization

BN was introduced in [40]↑to enable faster training of deep neural networks by preventing saturated gradients. This is achieved via normalization of inputs before each non-linear activation layer. In a BN layer, each batch x_B is normalized as shown in Eqn. 4.1.

$$BN(x_B) = \gamma \times \frac{x_B - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \tag{4.1}$$

During training, μ_B and σ_B^2 are the mean and variance of x_B. At test time, they are the estimated population mean and variance, as approximated by a moving average over training batches. γ, β are learnable parameters that allow the network to undo the normalization, if required. ϵ is a small additive constant in the denominator for numerical stability.

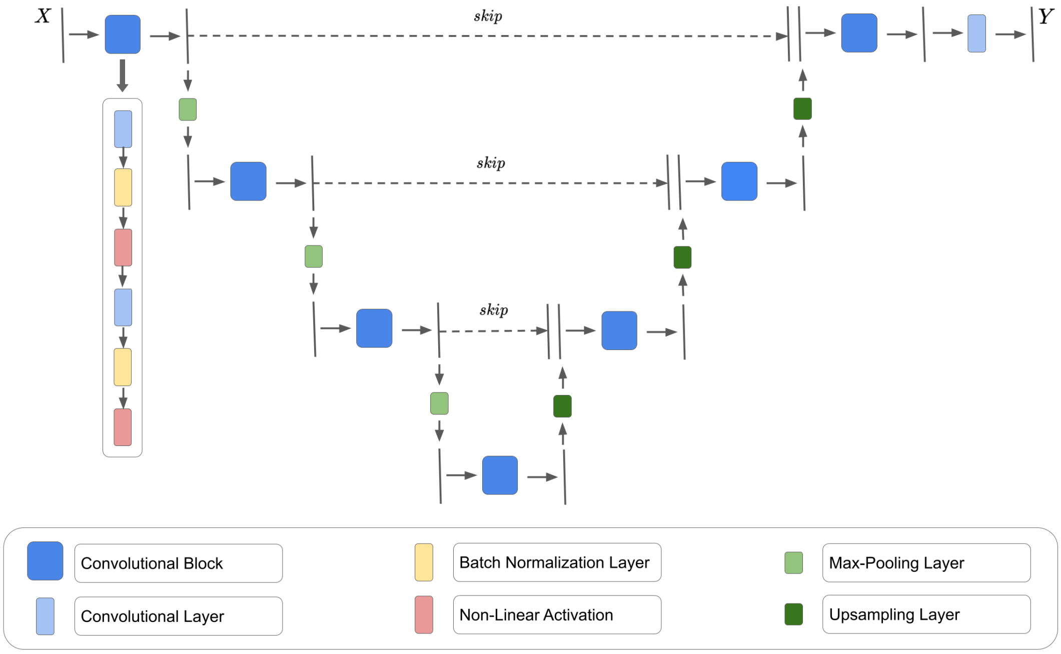


Figure 4.1: Representative CNN architecture for image segmentation. In the proposed method, the convolutional layers (light blue) are shared across distributions, while the batch normalization layers (yellow) are specific to each distribution.

4.2.2 Common CNN Architectures for Image Segmentation

A representative CNN architecture for image segmentation is shown in Fig. 4.1. It is based on the widely used U-Net architecture [205]↑. The key idea of the proposed method is to capture distribution-agnostic semantic information via shared convolutional layers, and distribution-specific image statistics in BN layers. Therefore, the method requires BN layers (Sec. 4.2.1) to be present in the network architecture. In Fig. 4.1, a BN layer exists between every convolutional layer (except the last one) and the application of non-linear activation.

We emphasize, however, that the proposed method is agnostic to the specifics of the chosen CNN architecture. In principle, it should be compatible with recent architectural improvements, including dense connections [206]↑, attention modules [207]↑, transformers [118]↑, among others.

4.3 Method

4.3.1 Splitting Parameters into Distribution-Agnostic and Distribution-Specific

According to the notation introduced in Sec. 2.1, the segmentation CNN is T_Θ. It maps the input images X to predicted segmentations Y. Let us split the set of all learnable parameters, Θ, into two subsets: (1) distribution-agnostic convolutional parameters, θ, and (2) distribution-specific batch normalization parameters, ϕ. Specifically, ϕ includes μ_B, σ_B^2, γ and β of all the BN layers.

4.3.2 Supervised Learning on Training Distributions

We assume access to labelled training datasets $\mathcal{D}^L_{tr,1}, \mathcal{D}^L_{tr,2} \cdots \mathcal{D}^L_{tr,K}$ from K training distributions. Let ϕ_k be the distribution-specific parameters of training distribution k. During training, each batch consists of data from only one distribution, with all training distributions covered successively. In a training iteration when the batch consists of data from distribution k, the parameters θ and ϕ_k are updated via stochastic gradient descent, while the parameters $\phi_{k'}$ for $k' \neq k$ are frozen.

4.3.3 Estimating the Closest Training Distribution to the Given Test Distribution

Consider a test distribution $K+1$ with a small labelled dataset $\mathcal{D}^L_{ts,K+1}$. We split this small dataset into two halves: one for training, $\mathcal{D}^L_{tr,K+1}$ and the other for validation, $\mathcal{D}^L_{vl,K+1}$. We evaluate the performance of T_Θ on $\mathcal{D}^L_{tr,K+1}$, using each of the subsets ϕ_k, $k = 1, 2, \cdots K$ one at a time. If ϕ_{k^*} leads to the best accuracy, we infer that among the already learned distributions, distribution k^* is the closest to given test distribution $K+1$.

4.3.4 Transfer Learning on the Given Test Distribution

Keeping shared parameters θ fixed, an additional set of BN parameters ϕ_{K+1} is initialized with ϕ_{k^*} and fine-tuned using $\mathcal{D}^L_{tr,K+1}$ with standard stochastic gradient descent minimization. The optimization is stopped when the performance on the validation set $\mathcal{D}^L_{vl,K+1}$ stops improving. Now, the network T_Θ can segment data from all distributions $k = 1, 2, \ldots K, K+1$ using their respective distribution-specific parameters, ϕ_k.

4.3.5 Favourable Properties for Lifelong Learning

In the spirit of lifelong learning, the proposed approach allows learning on new distributions with only a few labelled examples. This is enabled by utilizing the knowledge obtained from learning on the old distributions, in the form of the trained distribution-agnostic parameters.

The fact that the number of distribution-specific parameters is small comes with two advantages. One, that they can be tuned for a new distribution by training with a few labelled images quickly and with minimal risk of overfitting. Second, they can be saved for each distribution without significant memory footprint.

Finally, catastrophic forgetting [82]↑is a key problem in lifelong learning. It refers to drastic degradation in performance on previously learned distributions, when the model is adapted for shifted distributions. In the proposed approach, this problem does not arise by construction, because of the explicit separate modeling of shared and private parameters.

4.4 Experiments and Results

4.4.1 Datasets

We use images from 4 publicly available datasets: Human Connectome Project (HCP) [186]↑, Alzheimers Disease Neuroimaging Initiative (ADNI) [188]↑↑, Information eXtraction from Images (IXI)↑ and Autism Brain Imaging Data Exchange (ABIDE) [187]. HCP provides both T1-weighted and T2-weighted images. ADNI consists of T1-weighted images. From the ABIDE dataset, we use T1-weighted images from the center AC. Finally, we use T2-weighted images from the IXI dataset. (Please refer to Sec. 3.1 for a detailed description of the differences between these datasets).

We treat each dataset as a distinct distribution. The IDs assigned to each distribution and the number of images available in each dataset are shown in Table 4.1. We treat distributions $k = 1, 2, 3$ as initially available training distributions, and $k = 4, 5$ as new test distributions that are encountered after the initial training. As mentioned in Sec. 2.1, N_{tr} and N_{ts} indicate the number of training and test images. In this chapter, we introduce an additional variable, $N_{tr}^{scratch}$; its meaning is explained in Sec. 4.4.4.

As mentioned Sec. 3.1, these datasets do not provide manual annotations; the ground truth annotations are instead generated using FreeSurfer [194]↑. FreeSurfer takes approximately 10 hours to generate the segmentation for one volumetric image. Due to this constraint, we generated segmentations for roughly 50 images for HCP and ADNI, even though those datasets contain a much higher number of images. Nevertheless, we find that even with around 30 images, a segmentation CNN can be trained to provide satisfactory performance.

Distribution ID k	Train / Test	Center ID	MR Modality	N_{tr}	$N_{tr}^{scratch}$	N_{ts}
1	Train	HCP-T1	T1w	30	30	20
2	Train	HCP-T2	T2w	30	30	20
3	Train	ADNI-T1	T1w	30	30	20
4	Test	ABIDE-AC-T1	T1w	4	30	20
5	Test	IXI-T2	T2w	4	30	20

Table 4.1: Datasets for evaluation of the proposed transfer learning method.

4.4.2 Pre-processing

Image normalization

For each image volume, the intensities are normalized by dividing by their 98th percentile.

No other pre-processing steps were carried out for the experiments done in this project. Later, we realized that the extent of distribution-specific effects can be reduced by using a more extensive pre-processing pipeline. Thus, for later projects, we additionally included steps such as bias field correction, skull stripping (for brain imaging datasets) and spatial resolution matching. Please see Sec. 5.4.2 and Sec. 6.4.2 for more details.

4.4.3 Common Implementation Details for all Experiments

Network architecture

While the domain-specific BN layers can be incorporated in any standard CNN, we work with the widely used U-Net [205] architecture with minor alterations. Namely, our network has a reduced depth with three max-pooling layers and a reduced number of kernels: 32,64,128,256 in the convolutional blocks on the contracting path and 128,64,32 on the upscaling path. Also, bilinear interpolation is preferred to deconvolutional layers for upscaling in view of potential checkerboard artifacts [208].

Training details

The network is trained to minimize the dice loss [209] to reduce sensitivity to imbalanced classes. The initial network trains in about 6 hours, while the domain-specific BN modules can be updated for a new domain in about 1 hour, on a Nvidia Titan Xp GPU.

Evaluation Metric

Evaluation of segmentation accuracy is done via on Dice score, averaged over all foreground labels and over N_{ts} volumetric images from the appropriate distribution (see Table 4.1).

4.4.4 List of Experiments and Specific Implementation Details

(I) Baseline and Benchmark

For each distribution, we train an independent segmentation CNN, using a dataset of that distribution with $N_{tr}^{scratch}$ labelled images. For the training distributions ($k = 1, 2, 3$), the accuracy provided by the independent networks serves as a baseline that the other CNNs with shared parameters must preserve. For the test distributions distributions ($k = 4, 5$), the performance of the independent networks is the benchmark that we seek to achieve via transfer learning. For transfer learning, we use much fewer labelled samples (N_{tr}) from the test distributions, and using the knowledge of the previously learned distributions via the fixed shared parameters, θ.

(II) Transfer Learning by Adapting ϕ

We implement the proposed transfer learning approach in three steps: First, we jointly train a CNN on training distributions ($k = 1, 2, 3$) with shared convolutional parameters, θ, and distribution-specific batch normalization parameters, ϕ. Second, given a new test distribution, we determine the closest training distribution, as described in Sec 4.3.3. Finally, the distribution-specific parameters ϕ are initialized as those of the closest training distribution, and fine-tuned for the test distribution, as described in Sec 4.3.4.

(III) Transfer Learning by Adapting ϕ, θ

In order to analyze the importance of the parameter splitting proposed in this work, we investigate what happens if all the CNN parameters are updated for each new test distribution. To test this, we first jointly train a CNN, with all parameters shared, including those of the BN layers. In contrast to the training regime of described in Sec. 4.3.2, each training batch in this experiment contains randomly chosen images from all training distributions. This ensures that the shared BN parameters can be tuned well for all training distributions. For test distributions, we check if histogram equalization [210]↑can ensure good performance with the shared CNN. Finally, we check if performance improvement on the test distributions can be obtained if all the CNN parameters (θ, ϕ) are fine-tuned with N_{tr} images of the new distribution. In this setting, we also check if the updated parameters can retain performance on the initial training distributions.

4.4.5 Results

(I) Baseline and Benchmark

Quantitative results of baseline and benchmark experiments are shown in top five rows of Table 4.2. Along with the average Dice score over all 14 foreground labels, individual Dice scores of some important tissues are shown as well. It can been observed that using labelled datasets from each distribution, an independent CNN can provide highly accurate results for distributions, $k = 1, 2, 3, 4$. For $k = 5$, the segmentation performance of the individual CNN is substantially lower than the other distributions. This may indicate inherent segmentation ambiguity in this dataset.

Train	θ	ϕ	Test	Thal	Hipp	Amyg	Ventr	Caud	Puta	Pall	AVG.
Baseline and Benchmark											
1	θ_1	ϕ_1	1	0.92	0.86	0.85	0.90	0.90	0.89	0.75	0.87
2	θ_2	ϕ_2	2	0.91	0.84	0.84	0.89	0.89	0.88	0.74	0.85
3	θ_3	ϕ_3	3	0.91	0.87	0.81	0.94	0.86	0.88	0.85	0.88
4	θ_4	ϕ_4	4	0.92	0.88	0.85	0.93	0.91	0.90	0.85	0.89
5	θ_5	ϕ_5	5	0.88	0.79	0.77	0.80	0.79	0.82	0.79	0.81
Joint Learning of Distributions $1, 2, 3$ with Distribution-Specific BN Parameters											
$1, 2, 3$	θ_{123}	ϕ_1	1	0.91	0.85	0.84	0.89	0.89	0.88	0.73	0.86
$1, 2, 3$	θ_{123}	ϕ_2	2	0.91	0.85	0.84	0.89	0.88	0.87	0.75	0.86
$1, 2, 3$	θ_{123}	ϕ_3	3	0.91	0.87	0.82	0.94	0.87	0.88	0.85	0.88
Transfer Learning by Adapting ϕ for Distribution 4											
$1, 2, 3$	θ_{123}	ϕ_1	4	0.62	0.29	0.22	0.17	0.68	0.58	0.46	0.43
$1, 2, 3$	θ_{123}	ϕ_2	4	0.16	0.00	0.00	0.00	0.04	0.02	0.00	0.03
$1, 2, 3$	θ_{123}	ϕ_3	4	0.72	0.27	0.31	0.55	0.57	0.52	0.30	0.46
$1, 2, 3$	θ_{123}	$\phi_{3\rightarrow4}$	4	0.88	0.83	0.77	0.91	0.88	0.85	0.77	0.84
Transfer Learning by Adapting ϕ for Distribution 5											
$1, 2, 3$	θ_{123}	ϕ_1	5	0.00	0.02	0.06	0.01	0.00	0.00	0.00	0.01
$1, 2, 3$	θ_{123}	ϕ_2	5	0.35	0.12	0.27	0.23	0.41	0.28	0.37	0.29
$1, 2, 3$	θ_{123}	ϕ_3	5	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01
$1, 2, 3$	θ_{123}	$\phi_{2\rightarrow5}$	5	0.77	0.69	0.69	0.76	0.67	0.71	0.71	0.72

Table 4.2: TL-BN Quantitative Results (DICE). As summarized in Table 4.1, the distribution IDs 1, 2, 3, 4 and 5 stand for HCP-T1, HCP-T2, ADNI-T1, ABIDE-AC-T1 and IXI-T2, respectively.

(II) Transfer Learning by Adapting ϕ

Quantitative results of the proposed method are shown in the lower 11 rows of Table 4.2. Firstly, it can be observed that a joint CNN trained with shared convolutional parameters, but shared BN parameters provides the same performance for the training distributions $k = 1, 2, 3$ as the independent CNNs for each distribution.

For a new distribution $k = 4$, using the BN parameters ϕ_3 leads to the best performance. Thus, we infer that the training distribution $k = 3$ is the closest to $k = 4$. Starting with ϕ_3 and fine-tuning the BN parameters for $k = 4$, we find that the Dice scores for all the structures improve dramatically and are comparable to the benchmark performance of independent CNN trained exclusively for $k = 4$. Furthermore, as the original ϕ_k for k=1,2,3 are saved, the performance on the training distributions ($k = 1, 2, 3$) is identically preserved even after the transfer learning procedure. Similar results can be seen for the other new distribution, $k = 5$. In this case, the training distribution $k = 2$ is the closest to $k = 5$.

The improvement in the segmentations for new distributions after fine-tuning the BN parameters can also be observed qualitatively in Fig. 4.2.

(III) Transfer Learning by Adapting ϕ, θ

Quantitative results of this analysis experiment are shown in Table 4.3. First, we observe that even with all shared parameters, a single CNN can learn to segment images of multiple distributions, provided sufficient training data is available from all the distributions at once. However, its performance severely degrades for unseen distributions.

Histogram equalization (denoted by HE) to the closest distribution is unable to improve performance significantly. As well, fine-tuning all the parameters for the test distribution causes the CNN to catastrophically forget [82]↑the learning of the training distributions - that is, the performance on the training distributions severely degrades.

4.5 Discussion

In this chapter, we presented a transfer learning approach for improving robustness of a segmentation CNN to acquisition-related distribution shifts in medical imaging. The proposed method enables a segmentation CNN to be trained on an initial set of scanners and imaging protocols, and further adapted to new scanners or protocols with only a few labelled images and without degrading performance on the previous scanners. This was achieved by learning batch normalization parameters for each scanner, while sharing the con-

Train	θ	ϕ	Test	Thal	Hipp	Amyg	Ventr	Caud	Puta	Pall	AVG.
Joint Learning of Distributions $1, 2, 3$ with no Distribution-Specific Parameters											
$1, 2, 3$	θ_{123}	ϕ_{123}	1	0.91	0.85	0.82	0.89	0.88	0.88	0.75	0.85
$1, 2, 3$	θ_{123}	ϕ_{123}	1	0.89	0.84	0.82	0.88	0.86	0.86	0.70	0.83
$1, 2, 3$	θ_{123}	ϕ_{123}	1	0.91	0.85	0.79	0.94	0.86	0.87	0.83	0.86
Transfer Learning by Adapting θ, ϕ for Distribution 4											
$1, 2, 3$	θ_{123}	ϕ_{123}	4	0.75	0.25	0.06	0.79	0.43	0.32	0.07	0.38
$1, 2, 3$	θ_{123}	ϕ_{123}	4 (HE)	0.64	0.43	0.18	0.75	0.63	0.58	0.30	0.50
$1, 2, 3$	$\theta_{123 \to 4}$	$\phi_{123 \to 4}$	4	0.91	0.86	0.74	0.92	0.89	0.86	0.79	0.85
Effect of Transfer Learning on Initial Training Distributions											
$1, 2, 3$	$\theta_{123 \to 4}$	$\phi_{123 \to 4}$	1	0.87	0.81	0.77	0.87	0.86	0.72	0.67	0.80
$1, 2, 3$	$\theta_{123 \to 4}$	$\phi_{123 \to 4}$	2	0.67	0.42	0.51	0.11	0.64	0.40	0.41	0.45
$1, 2, 3$	$\theta_{123 \to 4}$	$\phi_{123 \to 4}$	3	0.80	0.76	0.65	0.75	0.73	0.72	0.77	0.74
Transfer Learning by Adapting θ, ϕ for Distribution 5											
$1, 2, 3$	θ_{123}	ϕ_{123}	5	0.42	0.18	0.18	0.44	0.27	0.20	0.03	0.24
$1, 2, 3$	θ_{123}	ϕ_{123}	5 (HE)	0.29	0.14	0.16	0.44	0.26	0.29	0.01	0.23
$1, 2, 3$	$\theta_{123 \to 5}$	$\phi_{123 \to 5}$	5	0.86	0.78	0.76	0.80	0.76	0.80	0.74	0.79
Effect of Transfer Learning on Initial Training Distributions											
$1, 2, 3$	$\theta_{123 \to 5}$	$\phi_{123 \to 5}$	1	0.27	0.02	0.17	0.00	0.05	0.00	0.00	0.08
$1, 2, 3$	$\theta_{123 \to 5}$	$\phi_{123 \to 5}$	2	0.57	0.57	0.56	0.74	0.66	0.52	0.53	0.59
$1, 2, 3$	$\theta_{123 \to 5}$	$\phi_{123 \to 5}$	3	0.15	0.03	0.16	0.01	0.11	0.04	0.00	0.07

Table 4.3: TL-BN Analysis Experiments (DICE). HE stands for histogram equalization.

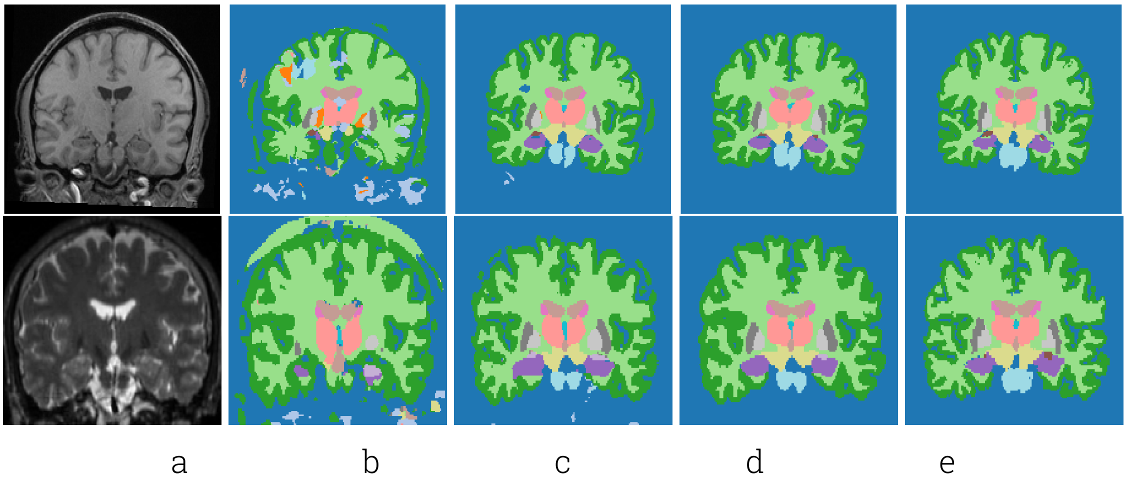


Figure 4.2: Qualitative results: (a) images from test distributions k, segmentations predicted by CNNs with parameters (b) θ_{123}, ϕ_{k^*} (baseline), (c) θ_{123}, $\phi_{k^* \to k}$ (adapted via the proposed transfer learning), (d) θ_k, ϕ_k (benchmark) and (e) ground truth annotations, with $\{k, k^*\}$ as {4, 3} (top) and {5, 2} (bottom).

volutional filters between all scanners. The method was evaluated on brain imaging datasets acquired from 5 combinations of scanners and imaging protocols.

While showing promising performance, the proposed method has the constraint of requiring an annotated dataset in the test distribution. In further chapters, we develop methods to alleviate this requirement. Secondly, the choice of adaptation parameters in this chapter was heuristically motivated. We improve upon this choice in the following chapters by proposing different adaptation subsets which are motivated by the domain knowledge of the causes of acquisition-related distribution shifts.

Chapter 5

A Test-Time Adaptation Approach for Robust Medical Image Segmentation

This chapter is based on the publication "Test-time adaptable neural networks for robust medical image segmentation." [67]↑. In this work, we tackle the lack of robustness of convolutional neural networks (CNNs) to acquisition-related distribution shifts (DS) in a test-time adaptation (TTA) setting. In contrast to transfer learning (chapter 4), TTA proposes to carry out model adaptation without requiring any labelled images from the test distribution.

5.1 Introduction

The TTA Setting

Standard supervised machine learning operates in a training-testing binary. That is, a model is first trained in a supervised manner, using labelled data from a training distribution. The model parameters are fixed thereafter, and directly used to make predictions for inputs from a test distribution. This leads to performance drop in the presence of DS. TTA deviates from such a strict training-testing demarcation. It argues that in the absence of knowledge about the shifted test distribution during training, it may be necessary to introduce some adaptability into the model. This can potentially enable the model to deal with images from the shifted distribution (for e.g., arising from new scanners and / or protocols). Thus, instead of fixing the model parameters after supervised learning on the training distribution, they are adapted to suit each test image. The adapted parameters are used to make predictions for that image.

Comparison of TTA with UDA and DG

Unsupervised domain adaptation (UDA) is a widely used setting for resolving DS. It requires the entire training dataset to be present while carrying out the adaptation for each new test distribution. We believe that this is a particularly stringent requirement in medical imaging, where sharing datasets across institutions often requires regulatory and privacy clearances.

Domain Generalization (DG), another well-studied setting, requires the access to the training dataset only for the initial training, and not during inference. A trained model is transported to the test site and used to perform inference, without requiring access to a labeled or unlabeled training set. This is clearly advantageous considering the aforementioned challenges in data sharing. As compared to sharing the training dataset across institutions, it is much easier to transport a trained model for usage with images from new test distributions. However, the lack of adaptation at test-time can lead to performance degradation if the encountered test distribution differs from all distributions seen during training.

TTA combines the advantages of UDA and DG. Like DG, TTA does not require access to any sample other than the test sample at hand. Yet, like UDA, it allows the model to be adapted to the test distribution before being used for prediction.

TTA achieves this best-of-both-worlds scenario by condensing the the information contained in the training dataset into a dedicated adaptation model. Such a model is transferred to the test site, instead of transferring the training dataset itself. The model is then used for facilitating the adaptation for each test image.

TTA for Robust Image Segmentation

We focus on the task of medical image segmentation. We design a segmentation CNN as a concatenation of two sub-networks: (a) a relatively shallow image normalization CNN, followed by (b) a deep CNN that segments the normalized image. We train both these sub-networks jointly in a supervised manner - using a training dataset, consisting of annotated images from the training distributions (a particular scanner and protocol setting). Then, at test time, we freeze the parameters of the deep segmentation sub-network, but adapt the image normalization sub-network for each test image. The adaptation is guided by an implicit prior on the predicted segmentation labels. More specifically, the adaptation requires that the predicted segmentation be plausible, according to the segmentations observed in the training dataset. For dictating such plausibility, we employ denoising autoencoders (DAEs) [211]↑.

Evaluation

We validate the proposed method on multi-center Magnetic Resonance Imaging (MRI) datasets of three anatomies - brain, heart and prostate. The proposed test-time adaptation consistently provides performance improvement, demonstrating the promise and generality of the approach.

Salient Characteristics of the Proposed Method

To the best of our knowledge, this was the first work in the literature to propose TTA for tackling the cross-scanner robustness problem in CNN-based medical image segmentation. We believe that the proposed inference time adaptation strategy has the following benefits.

- A normalization sub-network is used to adapt to each test image specifically. Thus, the method does not rely on similarity of the test image to previously seen training samples, as is the case in the majority of the works in the literature.
- The adaptable normalization sub-network is kept relatively shallow. This prevents it from introducing substantial structural change in the input image, while having sufficient flexibility to correct errors in the predict segmentation.
- We freeze the majority of the overall parameters at their pre-trained values (those of deep segmentation network). Thus, we retain the benefits of the initial supervised training, potentially done with a large number of training examples and, therefore, valuable for the segmentation task.
- The models used to drive the test-time adaptation, DAEs, can be very expressive. They can potentially exploit high-level cues such as context and shape in order to suggest corrections in the predicted segmentation.

5.2 Background

5.2.1 Denoising Autoencoders

Denoising Autoencoders (DAEs) [211]↑are neural networks with an encoder-decoder structure. They are trained to receive noisy / corrupted data as inputs and to output noise-free / clean / denoised data. Thus, their training dataset a set of pairs of corrupted and corresponding clean data points. Such a set is generated by corrupting clean data points with a known corruption process. A commonly used corruption process is one that adds Gaussian noise to clean data points. Furthermore, a common loss function for training the

DAE is the L_2 distance between the DAE's output and the known denoised data point. In this setting, the gradient of the clean data's probability density function can be expressed in terms of the DAE's reconstruction error [212]↑.

5.3 Method

5.3.1 Splitting Parameters into Image-Agnostic and Image-Specific

According to the notation introduced in Sec. 2.1, the segmentation CNN is T_Θ. It maps the input images X to predicted segmentations Y. We propose to train T_Θ in a supervised manner using labelled data from the training distribution, and further adapt it for each test image.

The first design question in this approach is: which parameters to adapt at test time? To answer this question, we note that DS due to changing imaging protocols and scanners manifest in the form of differences in low-level intensity statistics and contrast changes between different tissue types. Accordingly, we posit that a relatively shallow image-specific normalization sub-network might provide sufficient adaptability to obtain accurate segmentations within the relevant DS.

With this reasoning, we propose to formulate $T_\Theta(X)$ as $S_\theta(N_\phi(X))$ - a concatenation of two sub-networks. N_ϕ is a shallow normalization CNN - it takes as input the image to be segmented and outputs $X_n = N_\phi(X)$, X_n being a normalized image. S_θ is a deep segmentation CNN - it takes as input the normalized image, and outputs the predicted segmentation, $Y = S_\theta(X_n)$. The parameters ϕ are image-specific (adapted for each image), while the parameters θ are image-agnostic (shared across all training and test images). As described in Sec. 5.3.2, both ϕ and θ are learned in a supervised manner using a labelled dataset in the training distribution. Thereafter, θ are fixed, while ϕ are adapted for each test image.

Architecture of the Normalization Sub-Network

We model N_ϕ as a residual CNN. It processes the input image with n_N convolutional layers, each with kernel size k_N and stride 1. We employ no spatial down-sampling or up-sampling in N_ϕ and have it output the same number of channels as the input image. The chosen architecture for N_ϕ has the following advantages.

- Sufficient flexibility to model contrast transformations: We hypothesize that such an adaptable normalization module could enable an image-specific intensity transformation in order to alter the test image's contrast such that the pre-trained segmentation sub-network, $S_\theta(.)$, can accurately carry out the segmentation.

- Insufficient flexibility to cause structural changes: We restrict the kernel size (k_N) as well as the number of layers (n_N) to relatively small values. By doing so, we aim to limit N_ϕ to expressing intensity transformations that are sufficient for modeling contrast changes, but insufficient for substantially altering the image content by adding, removing or moving anatomical structures.
- Retaining benefits of supervised learning by freezing majority of the parameters: An important benefit of our formulation is that it freezes the majority of the overall parameters (those of the S_θ) at their pre-trained values. Thus, the method can leverage benefits of supervised learning by utilizing S_θ at its full capacity, as described in Sec. 5.3.2.

A representative schematic of the proposed parameter splitting is shown in Fig. 5.1. If the convolution kernel size k_N is set to 1, N_ϕ is strictly restricted to modeling intensity transformations without allowing for any structural changes. In our experiments, however, setting k_N to 1 led to training instabilities. Thus, we set k_N to 3, as described in Sec. 5.4.3. Further, note that the proposed architecture of N_ϕ is a design choice. An alternative choice could be to model N_ϕ as a parametric polynomial function of a certain degree. Finally, we emphasize that the proposed design of N_ϕ is suitable for tackling acquisition-related DS. For other types of DS, this design must be suitably adapted. See chapter 7 for an elaborate discussion.

5.3.2 Supervised Learning on Training Distribution

We assume that we have access to a training dataset $\mathcal{D}^L_{tr}$: $\{(x_i, y_i)|\ i = 1, 2, \ldots N\}$, where $x_i \sim P_{tr}(X)$ are sample images from a training distribution and y_i are corresponding ground truth segmentations. The $\mathcal{D}^L_{tr}$ can be composed of images coming from only one scanner and protocol setting or contain images from multiple scanners and protocols. The proposed method is agnostic to the formation of $\mathcal{D}^L_{tr}$. In our experiments, in view of potential difficulties in annotating and aggregating data from multiple imaging centers, we restrict the training distribution to a particular combination of imaging scanner and protocol setting. Given this annotated dataset, the goal is to provide an automatic segmentation method that works for new images sampled from not only the training distribution ($\mathcal{P}_{tr}(X)$), but also unseen test distributions ($\mathcal{P}_{ts}(X)$).

We train $T_\Theta(X) = S_\theta(N_\phi(X))$ using the labelled training dataset, $\mathcal{D}^L_{tr}$ via empirical risk minimization. Thus, the optimal parameters, $\{\theta^*, \phi^*\}$, are estimated

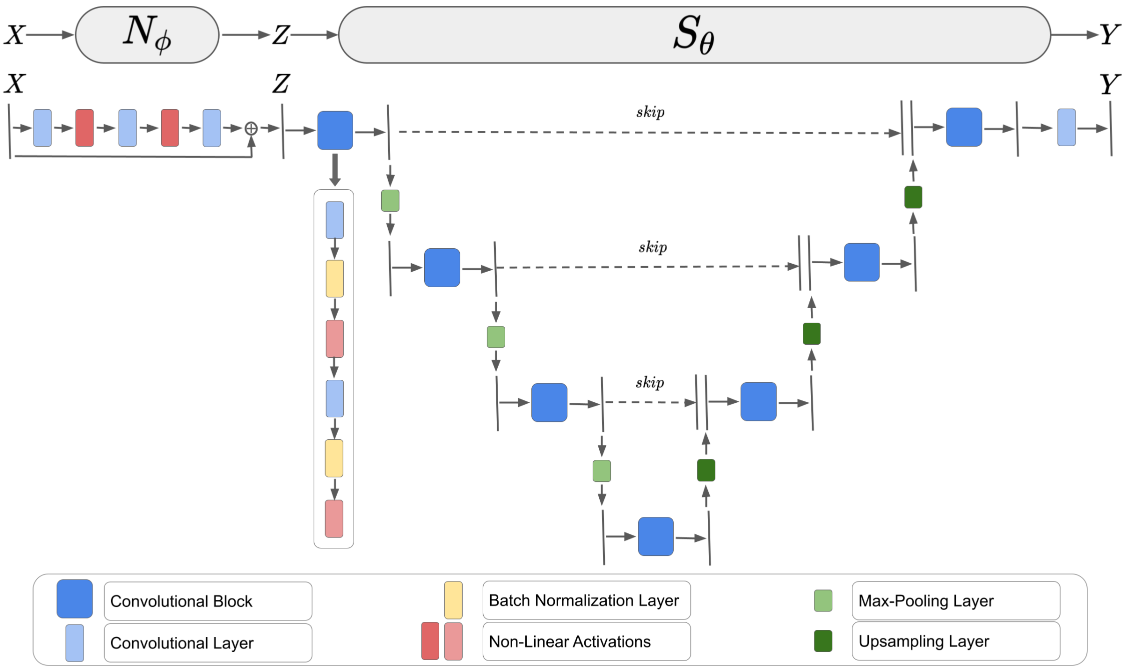


Figure 5.1: The Proposed parameter split for TTA. A relatively shallow normalization sub-network, N_ϕ, is image-specific, while a deep normalized-image-to-segmentation sub-network, S_θ, is shared across all training and test images. The proposed method is agnostic to the choice of S_θ - it can be chosen to be any well-performing segmentation CNN from the literature. The non-linear activation functions may potentially be differently chosen in N_ϕ and S_θ (indicated by different shades of red).

by minimizing a supervised loss function:

$$\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}} \sum_i \mathcal{L}_{task}(S_\theta(N_\phi(x_i)), y_i) \tag{5.1}$$

where $\{x_i, y_i\}$ are image-label pairs from $\mathcal{D}^L_{tr}$, the sum is over all such pairs used for training and $\mathcal{L}_{task}$ is a loss function that measures dissimilarity between the ground truth labels and predictions of the network.

In non-adaptable networks, the common application of CNNs for segmentation, once the optimal parameters are estimated, the segmentation for a new image x is obtained as $y^* = S_{\theta^*}(N_{\phi^*}(x))$. In this work, we modify this procedure by introducing an adaptation step at test time.

5.3.3 How to Drive Adaptation at Test-Time?

The optimization in Eq. 5.1 depends on the training dataset, and in particular, on the intensity statistics of the training images, $x_i \sim P_{tr}(X)$. Thus, if confronted with shifts in the input distribution at test time, the pre-trained mapping $S_{\theta^*}(N_{\phi^*}(.))$ may not be reliable. To address this, we propose to use

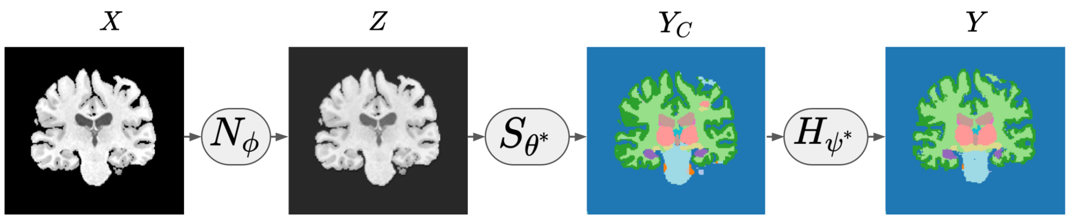


Figure 5.2: Workflow of the proposed method: For each test image, N_ϕ is adapted such that the resulting segmentation is plausible, as gauged by a DAE, H_{ψ^*}. The star superscript next to the parameters indicates that the models S_{θ^*} and D_{ψ^*} are learned on the training distribution and fixed thereafter, while the model N_ϕ is adapted for each test image. X is the input image, Z is the normalized image, Y_C is the predicted segmentation that potentially may be corrupted due to the DS problem and Y is the denoised segmentation outputted by the DAE. TTA is driven by adapting N_ϕ to make Y_C like Y.

the pre-trained parameters as an initial estimate, further adapting them for each test image. In order to implement this idea, the first design question was which parameters to update at test-time? We answered this in Sec. 5.3.1. Now, we turn to second important design question: how to drive the adaptation at test time, without label information and with only the test image available?

Driving TTA by Increasing Plausibility of Predicted Segmentations

The main challenge in TTA is the lack of label information and additional images. The model only has access to the test image to which it should adapt. In this scenario, we drive the adaptation by requiring that the predicted segmentations be plausible, that is, similar to those seen in the training dataset. The underlying assumption here is that the DS in question pertain only to scanner and protocol changes, with the images otherwise containing similar structures, whether healthy or abnormal, as the training dataset.

Gauging Plausibility via Denoising Autoencoders

We use denoising autoencoders (DAEs) [211]↑to assess the similarity of a given segmentation to those in the training dataset. The idea is that if the predicted segmentation is implausible, the DAE will see it as a "noisy" segmentation and "denoise" it to produce a corresponding plausible segmentation. The output of the DAE can then be used as a pseudo-ground-truth segmentation to drive the TTA. Crucially, DAEs can be highly expressive - they have the capacity to leverage high-level cues, such as long-range spatial context and shape, in order to suggest corrections in predicted segmentations.

TTA Workflow

The workflow of our test-time adaptation method is depicted in Fig. 5.2. We leverage the available ground truth segmentations in the training dataset, $\mathcal{D}_{tr}^{L}$, to train a DAE, H_{ψ^*} [1]. H_{ψ^*} maps corrupted segmentations Y_c (which are not necessarily similar to those in the training dataset), to "denoised" segmentations Y (similar to those in the training dataset). The details of this training are explained in Sec. 5.3.4. For the time being, let us assume that we have a trained DAE, H_{ψ^*}. For a given test image x and a set of parameters for the normalization sub-network, ϕ, we treat the segmentation predicted by $S_{\theta^*}(N_\phi(x))$ as a "noisy" or "corrupted" segmentation. We pass this noisy segmentation through H_{ψ^*} and obtain its denoised version. Now, we update the parameters of N_ϕ so as to pull the predicted segmentation closer to its denoised version:

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}}\, \mathcal{L}_{task}(y_c, H_{\psi^*}(y_c));\ y_c = S_{\theta^*}(N_\phi(x)), \tag{5.2}$$

where $\mathcal{L}_{task}$ is a similar loss to that in Equation 5.1. Eqn. 5.2 denotes the test-time adaptation that we carry out for each test image x. This optimization is done iteratively (using either gradient descent or a variant thereof).

Expected Evolution of Predicted Segmentations during TTA

For a test image that is not from the training distribution, the predicted segmentation is likely to be corrupted at the beginning of TTA. For instance, it may appear similar to the one shown on the bottom-right in Fig. 5.2. The DAE takes this prediction as input and proposes a corrected segmentation, such as the one shown on the bottom-left in Fig. 5.2. Now, the parameters of N_ϕ are updated so as to minimize the dissimilarity between the DAE input and output. As TTA proceeds, the predicted segmentation, $S_{\theta^*}(N_\phi(x))$, becomes increasingly plausible, that is, similar to those in training dataset. Therefore, the DAE input and output become similar, resulting in small loss values and convergence of the TTA. Importantly, the adaptable normalization sub-network, N_ϕ, is relatively shallow and has a relatively small receptive field. Thus, the adaptation is free to change the contrast of the input image, but cannot introduce large structure alterations.

TTA Model Selection Criterion

The optimization runs for a pre-specified number of iterations and the optimal image-specific parameters $\hat{\phi}$ are chosen as the ones that provide the least dissimilarity between the DAE input y_c and output $H_{\psi^*}(y_c)$ during the iterations. We believe that this principled stopping criterion for TTA - rather than running the optimization for an arbitrary number of iterations - is an important characteristic of the proposed method.

The final segmentation is predicted as $\hat{y} = S_{\theta^*}(N_{\hat{\phi}}(x))$.

---

[1] We denote the DAE with H to indicate that it is a helper module that facilitates TTA. In the next chapter, we discuss other possible helper modules.

5.3.4 DAE Training

The DAE drives the proposed TTA, and is thus a key component of the proposed method. We model it as a 3D CNN. This potentially allows for learning of information about relative locations of different anatomical structures across volumetric segmentations as well as about their shapes in their entirety.

In order to train such a DAE, we generate a training dataset of pairs (y_i, y_{ci}), with $y_i \sim P(Y)$ (available from $\mathcal{D}_{tr}^{L}$) and $y_{ci} \sim P(Y_c|Y = y_i; \omega)$, a corruption process that we define in order to generate corrupted segmentations Y_c given clean segmentations Y. With this dataset, we train the DAE to predict $Y = H_\psi(Y_c)$ by minimizing the following loss function to estimate the parameters ψ^*:

$$\psi^* = \underset{\psi}{\operatorname{argmax}} \, \mathbb{E}[\mathcal{L}_{task}(H_\psi(Y_c), Y)] \tag{5.3}$$

Here, the expectation is over the joint distribution $P(Y, Y_c) = P(Y)P(Y_c|Y)$. Thus, we have

$$\psi^* = \underset{\psi}{\operatorname{argmin}} \sum_j \sum_i \mathcal{L}_{task}(H_\psi(y_{cij}), y_i) \tag{5.4}$$

where the index j denotes different samples obtained from $P(Y_c|Y = z_i; \omega)$, the outer sum is over the number of corrupted samples that we generate for each ground truth label y_i and $\mathcal{L}_{task}$ is a loss function that computes dissimilarity between the clean ground truth labels and the predictions of the DAE. As the DAE is trained in the space of segmentation labels, we use the same $\mathcal{L}_{task}$ as the one used for the initial supervised training (Eqn. 5.1) as well as for the TTA (Eqn. 5.2).

Noising strategy

The main design choice for the DAE training described above is the noising process, $P(Y_c|Y; \omega)$. This noising process is used to generate artificially degraded segmentations, simulating the inaccurate labels that the pre-trained CNN ($S_{\theta^*}(N_{\phi^*}(.))$) will likely predict when faced with input images from unseen test distributions. In this work, we follow a heuristic procedure for generating such noisy labels. We copy cubic patches from randomly chosen locations in the label image to other randomly chosen locations in the same image. In each training iteration of the DAE and for each clean label, the number of such patches (n_1) is sampled from an uniform distribution $U(0, n_1^{max})$. For each of these n_1 patches, its size (n_2) is sampled independently from another uniform distribution $U(0, n_2^{max})$. Thus, our noising process is defined by hyper-parameters: $\omega : \{n_1^{max}, n_2^{max}\}$. [2]

---

[2] If the noising process is chosen to be one that adds Gaussian noise to its inputs and if

5.3.5 Atlas initialization for TTA for large DS

DS robustness of the DAE

The DAE is used as a helper model to improve DS robustness of the segmentation CNN. Yet, the DAE itself is also modelled by a CNN and is trained in a self-supervised manner. Thus, the DAE could itself be vulnerable to DS in its inputs. That is, if the segmentations that are fed as inputs to the DAE during TTA are dissimilar to the DAE's training inputs, the outputs of the DAE may be unreliable.[3]

Assumption: DAE is robust to small DS

The DAE is immune to such unreliability so long as the probability distribution of the corrupted segmentations generated by our noising process (Sec.5.3.4) approximates that of the predictions of the pre-trained task CNN ($S_{\theta^*}(N_{\phi^*}(.))$) in response to test distribution images. For DS pertaining to scanner changes under the same imaging protocol, we assume that our noising process is able to satisfy this requirement. Specifically, for a given test image x, the segmentation predicted by the pre-trained task CNN as well as during the iterative TTA is $y_c = S_{\theta^*}(N_\phi(x))$. Now, if x is acquired using the same imaging modality and similar protocol as the training dataset images and has unknown ground truth segmentation y, then we assume that y_c can be seen as a corrupted segmentation that is a sample from our noising process $P(Y_c|Y = y)$.

DAE is unreliable when faced with large DS

The DAE's outputs can no longer be relied upon, however, when the training and test distribution images are very different. This is the case, for instance, when DS is caused via acquisition using different modalities or very different protocols, such as using MR for one image and CT for the other or T1-weighted MR for one and T2-weighted for the other. In such cases, the predictions of the pre-trained task CNN can be highly corrupted and may not be captured by the noising strategy described in Sec.5.3.4. Thus, the corresponding DAE outputs may no longer be reliable for driving TTA.

Atlas-based work-around for such scenarios

---

the DAE is trained by minimizing the L_2 loss (i.e if $L(H_\psi(y_{cij}), y_i) = ||H_\psi(y_{cij}) - y_i||_2$), then the gradient of the label prior, $P(Y)$, can be expressed in terms of the DAE reconstruction error [212]↑. This allows for explicit prior maximization [213]↑. However, this result does not generalize to different data corruption models, such as the noising strategy used in this work. On the other hand, a simple noising model that adds Gaussian noise is unlikely to mimic the inaccurate segmentations predicted by a pre-trained CNN in the face of acquisition-related DS in medical imaging.

[3] Indeed, this characteristic of the proposed method is one of its weaknesses. We tackle this weakness in the next chapter.

To deal with large DS consisting of imaging protocol changes, we utilize an affinely registered atlas, A, to first draw the predicted segmentations to a reasonable starting point from where the DAE can take over. Specifically, instead of directly carrying out the optimization as described in Eq. 5.2, we switch between minimizing $\mathcal{L}_{task}(y_c, H_{\psi^*}(y_c))$ and $\mathcal{L}_{task}(y_c, A)$, both with respect to ϕ. Here, $y_c = S_{\theta^*}(N_\phi(x))$ are the predictions of task CNN at any point during the iterative TTA.

Switching from the Atlas back to the DAE

We employ a threshold-based approach to decide when to switch from using the Atlas to using the DAE predictions for driving TTA. If $d(y_c, H_{\psi^*}(y_c))/d(y_c, A) \geq \alpha$ and $d(y_c, A) \geq \beta$, then we minimize $\mathcal{L}_{task}(y_c, H_{\psi^*}(y_c))$. Else, we minimize $\mathcal{L}_{task}(y_c, A)$. Here, d is a similarity measure between segmentations and α, β are hyper-parameters. In our experiments, we use the Dice loss [209] as $\mathcal{L}_{task}$ and the Dice score as d, with $\mathcal{L}_{task} = 1 - d$.

Justification for Threshold-based switching

The reasoning for the threshold-based switching is as follows: In the initial steps of the TTA (when dealing with large DS), the predicted segmentations will likely be extremely corrupted. Therefore, we would like to use the affinely registered atlas for driving the adaptation. Once the predicted segmentations improve and can be considered as samples from our noising process, we would like to switch to using the DAE outputs for driving the adaptation, as it has more flexibility than an affinely registered atlas. Our threshold-based switching procedure encodes two signals indicating improvement in the predicted segmentations: increased similarity between (1) DAE input and output (note that when the predicted segmentation is plausible according to the DAE, the DAE models an identity transformation) and (2) predicted segmentation and the atlas.

5.3.6 Integrating 2D Segmentation CNN with 3D DAE

As noted in Sec. 5.3.4, we model the DAE as a 3D CNN in order to leverage volumetric anatomical information. On the other hand, the CNN-based image segmentation literature is dominated by 2D CNN designs, mainly because 3D CNNs are hindered by memory issues and 2D CNNs already provide state-of-the-art segmentation performance in many cases [214]↑. In order for the proposed TTA method to be applicable to both 3D as well as 2D segmentation CNNs, we propose the following strategy for using 2D segmentation CNNs with our 3D DAE. In this case, the normalization sub-network, N_ϕ, is also a 2D CNN and we iteratively carry out the following two steps for T updates of the parameters of N_ϕ for a given 3D test image:

1. Predict the current segmentation for the entire 3D test image, by passing it through the 2D segmentation CNN in batches consisting of successive slices. Following this, pass the 3D predicted segmentation though the trained DAE to obtain its denoised version.

2. Initialize gradients with respect to ϕ to zero. Process the 3D test image in 2D batches consisting of successive slices as in step 1: For each batch, predict its segmentation, compute loss between the prediction and the corresponding batch of the denoised labels computed in step 1, and maintain a running sum of gradients of the loss with respect to ϕ. At the end of all batches, average gradients over the number of batches and update ϕ.

For large DS (that is, those including a change in imaging protocol), we use the threshold-based method described in Sec. 5.3.5 to determine whether to use the atlas or the DAE outputs as target labels for driving the adaptation in step 1. Furthermore, to save computation time, we update the denoised labels for the adaptation, i.e. run step 1, after f runs of step 2, instead of after every run.

5.4 Experiments and Results

5.4.1 Datasets

We validate the proposed method on multiple MRI datasets from three anatomies: brain, heart and prostate. Here, we describe which subsets of the datasets described in Sec. 3.1 were used for validation of the proposed method. Please refer to Sec. 3.1 for a detailed description of the datasets.

Brain MRI

We use images from 2 publicly available datasets: Human Connectome Project (HCP) [186]↑and Autism Brain Imaging Data Exchange (ABIDE) [187]↑↑↑↑. In the HCP dataset, both T1w and T2w images are available for each subject, while the ABIDE dataset consists of T1w images from several imaging sites. We use the HCP-T1 dataset as the training distribution, and the ABIDE (AC-T1) and HCP-T2 datasets as two independent test distributions. Being acquired with same modality, but from a different scanner, AC-T1 is deemed to be a small DS, while HCP-T2 is considered as a large DS.

Prostate MRI

We use the National Cancer Institute (NCI) dataset[181]↑as the training distribution. In particular, among the two sub-datasets within the NCI dataset, we use images from RUNMC institution. We use two test distributions to evaluate the proposed method: (i) the PROMISE12 dataset [42]↑and (ii) a private dataset from the University Hospital of Zurich (USZ) [182]↑. As described in Sec. 3.1, the PROMISE12 dataset consists of images from 4 distributions. During the evaluation of this method, we ignored this distinction - instead considering the entire set of images as one test distribution. For the NCI and USZ datasets, expert annotations are available for 3 labels for each image: background, central gland (CG) and peripheral zone (PZ), while the PROMISE12 dataset only provides expert annotations for the whole prostate gland (CG + PZ). Thus, we evaluate our predictions both for the whole gland as well as separate CG and PZ segmentations, respectively, for the different datasets.

Cardiac MRI

We use the Automated Cardiac Diagnosis Challenge (ACDC) dataset [43]↑as training distribution and the right ventricle segmentation challenge (RVSC) dataset [184]↑as the test distribution. or the ACDC dataset, annotations are available for LV (left ventricle) and RV (right ventricle) cavities and the LV myocardium. For the RVSC dataset, annotations are provided for the RV caity and the RV myocardium. Thus, we evaluate TTA performance based on RV cavity segmentation, the only structure that is common in both datasets, setting other predictions as background.

Table 5.1 shows our training, test and validation split (in terms of number of 3D images) for each dataset.

Anatomy	Dataset	Train / Test	N_{tr}	N_{vl}	N_{ts}
Brain	HCP-T1	Train	20	5	20
Brain	ABIDE-AC-T1	Test	10	5	20
Brain	HCP-T2	Test	20	5	20
Prostate	RUNMC	Train	15	5	10
Prostate	USZ	Test	28	20	20
Prostate	PROMISE12	Test	20	10	20
Heart	ACDC	Train	120	40	40
Heart	RVSC	Test	48	24	24

Table 5.1: Dataset details for the TTA-DAE experiments.

5.4.2 Pre-processing

We pre-process all images and segmentation labels with the following steps.

Bias Correction: Firstly, we remove any bias fields with the N4 algorithm [215].

Intensity Normalization: Secondly, we carry out $0-1$ intensity normalization per image as: $x_{normalized} = (x - x_p^1)/(x_p^{99} - x_p^1)$, where x_p^i denotes the i^{th}

percentile of the intensity values in the image volume, followed by clipping the intensities at 0 and 1.

Skull Stripping: For the brain datasets, this is followed by skull stripping, setting intensities of all non-brain voxels to 0.

Resolution matching: We train the segmentation CNN, T_{Θ}. in 2D due to GPU memory limitations, and to show applicability of the proposed method with widely used 2D architectures in the segmentation literature. We rescale all images to fixed pixel-size in the in-plane dimensions followed by cropping and / or padding with zeros to match the image sizes to a fixed size for each anatomy. The fixed pixel-sizes for the brain, prostate and cardiac datasets are 0.7mm^2, 0.625mm^2 and 1.33mm^2 respectively, while the fixed image size is 256x256 for all anatomies. The ground truth labels of the training and validation images are rescaled and cropped / padded in the same way as the corresponding images. Test images are also rescaled and cropped / padded before predicting their segmentations. The predicted segmentations, however, are rescaled back and evaluated in their original pixel-size to avoid any experimental biases.

We use a 3D CNN to model the DAE, H_{ψ}, as we believe that the incorporation of 3D organ structure can be vital for the DAE's performance. We preprocess the segmentation labels with rescaling and cropping / padding applied in all 3 dimensions. The fixed voxel-sizes are set to 2.8x0.7x0.7mm^3, 2.5x0.625x0.625mm^3 and 5.0x1.33x1.33mm^3 for the brain, prostate and cardiac datasets, respectively, while the fixed 3D image size is set to 64x256x256 for the brain images and 32x256x256 for the other two anatomies.

5.4.3 Common Implementation Details for all Experiments

Normalization Sub-Network, N_{ϕ}, architecture

We implement the normalization sub-network, N_{ϕ}, with $n_N = 3$ convolutional layers, with the respective number of output channels set to 16, 16 and 1, each using kernels of size $n_k = 3$. Keeping in mind the relatively small depth of N_{ϕ}, we equip it with an expressive activation function, $act(x) = exp(-x^2/\sigma^2)$, where the scale parameter σ is trainable and different for each output channel.

Segmentation Sub-Network, S_{θ}, architecture

For modeling the normalized-image-to-segmentation sub-network, S_{θ}, we use an encoder-decoder architecture with skip connections across corresponding depths, in spirit of the commonly used U-Net [205] architecture. Note that the proposed test-time adaptation strategy, the normalization sub-network and the DAE are agnostic to the architecture of the normalized-image-to-segmentation sub-network. Any architecture can be used instead of the U-Net that we used. Batch normalization [40]↑and the ReLU activation function [216] are used in S_θ as well as H_ψ. Bilinear upsampling is preferred to deconvolutions in light of the potential checkerboard artifacts while using the latter [208]↑.

Loss function, batch size, optimizer and learning rate

We use the Dice loss [209] as the loss function $\mathcal{L}_{task}$ in four cases: (i) supervised learning on the training distribution (Eqn. 5.1), (ii) DAE training (Eqn. 5.3), (iii) test-time adaptation (Eqn. 5.2) and (iv) atlas-based switching for large DS (Sec. 5.3.5). The batch size is set to 16 for the 2D segmentation CNN training and the test-time adaptation, and to 1 for the 3D DAE. We use the Adam optimizer [39] with default parameters and a learning rate of 0.001.

Evaluation Metrics

We evaluate the predicted segmentations by comparing them with corresponding ground truth segmentations using the Dice coefficient [217]↑and the 95th percentile of Hausdorff distance [218]↑. We report mean values of these scores computed in 3D, across foreground labels, all test images and across 3 runs of each experiment.

5.4.4 List of Experiments and Specific Implementation Details

(I) Baseline

For each anatomy, we train a segmentation CNN on the training distribution, and evaluate its performance on test images from the training as well as test distributions. We train T_Θ for 50000 iterations and chose the best models based on validation set performance. The performance of this CNN on the test distributions provides a baseline performance for the problem.

(II) Benchmark

We train specialized segmentation CNNs for each test distribution, using a separate training and validation set from that distribution. For the purposes of this work, the performance of such specialized CNNs forms the benchmark for the problem.

(III) Strong Baseline - Data Augmentation

In [101], data augmentation has been shown to be highly effective for improving cross-scanner robustness in medical image segmentation. Accordingly, we employ extensive data augmentations, consisting of geometric as well as intensity transformations. The geometric transformations are applied to both images as well as segmentations, while the intensity transformations are applied only to the images. Each transformation is applied with a probability of 0.25 to each image in a training mini-batch.

Intensity Transformations: As intensity transformations, we use gamma transformation ($x_{aug} = x^c; c \sim U(0.5, 2.0)$), brightness changes ($x_{aug} = x + b; b \sim U(0.0, 0.1)$) and additive Gaussian noise ($x^{ij}_{aug} = x^{ij} + n^{ij}; n^{ij} \sim N(0.0, 0.1)$, where the superscript ij is used to indicate that the noise is added independently for each pixel in the image).

Geometric Transformations: As geometric transformations, we use translation ($\sim U(-10, 10)$ pixels), rotation ($\sim U(-10, 10)$ degrees), scaling ($\sim U(0.9, 1.1)$) and random elastic deformations (obtained by generating random noise images between -1 and 1, smoothing them with a Gaussian filter with standard deviation 20 and scaling them with a factor of 1000) [219]↑. For the cardiac datasets, we observe that the images are acquired in different orientations, so for this anatomy, we add to the set of geometric transformations: rotations by multiple of 90 degrees and left-right and up-down flips.

As will be described in Sec. 5.4.5, such data augmentation provides substantial performance improvement, and is effective across anatomies. Due to its effectiveness, generality and ease of implementation, we treat this approach as a strong baseline, that we aim to improve upon with the proposed test-time adaptation.

(IV) Domain Generalization Methods

Several meta-learning based approaches [110]↑, [111]↑, [112]↑have been proposed for tackling the domain generalization problem. The main idea of such methods is to simulate the distribution shift problem during the training of the segmentation CNN. This is done by having meta-train and meta-test domains during training and requiring that the gradient updates for the meta-train distributions be such that the task loss is also minimized on the meta-test distributions. As we only had access to a single training distribution, we simulated meta-train and meta-test distributions by using different gamma transformations in each batch of training. Additionally, we used all other data augmentation transformations described in experiment (III).

(V) Post-Processing Methods

[143]↑propose post-processing the predicted segmentations with DAEs in order to increase their plausibility. We use our trained DAEs in order to carry out such post-processing on the segmentations predicted by the strong baseline.

(VI) Test-Time Adaptation

For the proposed method, we first trained the segmentation network on the training distribution along with data augmentation. Then, we adapted the normalization sub-network, N_ϕ, for each test image, according to the proposed framework.

DAE, H_ψ, architecture and training: We model the DAE, H_ψ, as a 3D CNN with an encoder-decoder architecture, as well as skip connections. We train the DAE for 50000 iterations and chose the best models based on validation set performance. We use data augmentation consisting of geometric transformations described in experiment (III), applied on the segmentations.

Noise Hyper-parameters for Generating DAE Training Data: We visually inspected the generated corrupted segmentations by using different noise hyper-parameters. Based on this, we chose the maximum number of patches to be copied, $n_1^{max} = 200$ and the maximum size of a patch, $n_2^{max} = 20$. During its training, we determined the best DAE model based on its denoising performance on a corrupted validation dataset, which we generate by corrupting each validation image 50 times with the noising process described in Sec. 5.3.4.

Number of adaptation iterations: For TTA for each test image, we run the inference-time optimization for $T = 500$ gradient updates for the brain datasets and for $T = 7500$ gradient updates for the other two anatomies (see Sec. 5.3.6, step 2). (This discrepancy is due to the differences in the number of slices of the datasets. In our implementation, each update of ϕ is performed with an average gradient over 16 batches for the brain datasets and over 2 batches for the prostate and cardiac datasets. To account for lower number of batches for the latter, we use larger number of gradient update steps. Thus, effectively, even with the different number of gradient updates, images from all datasets observe roughly the same number of batches during the optimization.) The denoised labels that are used to drive the optimization are updated every $f = 25$ steps (see Sec. 5.3.6, step 1).

TTA Model Selection: During the update iterations, parameters that lead to the highest Dice score between the DAE input and output are chosen as optimal for a given test image.

'Fast' TTA: Additionally, we run a separate 'fast' version of our method, where we carry out TTA with the aforementioned hyper-parameters for the first test image of each TD. For subsequent images of that TD, we initialize the parameters of the normalization module with the optimal parameters obtained for the first TD image. This provides a better starting point for the optimization, so we run it for $T = 100$ gradient updates for the brain datasets and for

$T = 1500$ gradient updates for the other anatomies. On a NVIDIA GeForce GTX TITAN X GPU, the test time adaptation requires about 1 hour for the first image of a particular TD and about 12 minutes for each image thereafter with our experimental implementation, which could be further optimized for time efficiency.

Hyper-parameters for atlas-based initial optimization for large DS: For the brain datasets, the SD consists of T1w images, while the TD$_2$ consists of T2w images, both from the HCP dataset, but from different subjects. In this case, we used the atlas based initial optimization described in Section 5.3.5. As the images in the HCP dataset are already rigidly registered, we create an atlas by converting the SD labels to one-hot representations and averaging them voxel-wise. To decide when the optimization switches from being driven by the atlas to the DAE, we use the thresholding-based method described in Sec. 5.3.5, setting the hyper-parameters $\alpha = 1.0$ and $\beta = 0.25$.

(VII) Unsupervised Domain Adaptation (UDA) Methods

UDA is widely proposed in the literature for tackling the domain shift problem. We compare the performance of the proposed TTA method with UDA works. We note that UDA methods work in a more relaxed setting, where the labelled training dataset is assumed to be available while adapting for each new test distribution. Although it may be challenging to meet this requirement in practice due to privacy concerns, we carry out this experiment to quantify the potential advantages of such approaches.

We conducted experiments with two representative UDA methods: 1) [88]↑, where an adversarial loss is employed to incentivize invariance in the SD and TD features, and 2) [90]↑, where a transformation network between the TD and the SD is trained, and then the transformed images are passed through the segmentation network.

Typically, UDA methods utilize a set of unlabelled images from the test distribution, $\mathcal{D}_{ts}^{UL,tr}$, for the adaptation, but use a separate set of test images from the TD, $\mathcal{D}_{ts}^{UL,ts}$, for evaluation. However, as ground truth labels of $\mathcal{D}_{ts}^{UL,tr}$ are not utilized for the adaptation, we present evaluations for this set as well, which was also proposed as a more suitable UDA setting in a recent work [220]↑. The size of the image sets $\mathcal{D}_{ts}^{UL,tr}$ and $\mathcal{D}_{ts}^{UL,ts}$ was the same as specified in Table 5.1.

(VIII) Ablation Studies

We conducted several experiments to analyze the importance of the design choices in the proposed method.

1. **Fast TTA**: We carried out a 'fast' version of our method, where the optimal normalization parameters, ϕ for the first test image of each test distribution

were used for initialization in the TTA for subsequent images of that distribution. After such improved initialization, the adaptation for subsequent images could be completed in fewer iterations, thus reducing the time required for the adaptation.

2. **Effect of adapting all parameters $\{\theta, \phi\}$ for TTA**: We studied the importance of restricting the adaptation to just the normalization sub-network. That is, we learned T_Θ on the training distribution, along with data augmentation, and adapted all its parameters, $\{\phi, \theta\}$, for each test image, according to the proposed framework.

3. **Expressiveness of N_ϕ**: We examined if the flexibility afforded by the adaptable normalization sub-network is sufficient for obtaining accurate segmentations via TTA. To this end, we learned T_Θ on the training distribution, along with data augmentation, and then adapted N_ϕ for each test image, driving the TTA using the ground truth labels of the test image. This is an ablation study that removed the DAE from the picture and asked the following question: if an oracle were available to drive the TTA, can N_ϕ be appropriately adapted to follow the oracle? This experiment was not done for the PROMISE and the RVSC datasets as the annotations for these datasets are for a different set of organs / tissues as compared to the corresponding training datasets.

4. **How does TTA compare to iterative post-processing?**: We asked the question: Is TTA required at all, or can the accuracy on test distributions be improved by simply passing the predicted segmentation multiple times through the DAE?

(IX) Convergence of TTA

As the convergence of TTA is not theoretically guaranteed, we empirically test if the adaptation at test time converges or not. To do so, we run the TTA for a large number of iterations, and track evolution of the segmentation performance.

5.4.5 Results

From the list of experiments (Sec. 5.4.4), the results of experiments (I) through (VI) are shown in Table 5.2.

(I) Baseline

The baseline results show substantial performance drop between the training and test distributions, showing the lack of robustness of CNNs to acquisition-related distribution shifts.

(II) Benchmark

Specialized CNNs for each test distribution provide substantially higher performance than the baseline. This indicates that learning is not inherently more difficult in the test distributions.

(III) Strong Baseline - Data Augmentation

A remarkable performance boost can be observed due to data augmentation, for cases where the training and test distribution images are acquired with the same imaging protocol, with different scanners. Nonetheless, there still remains a gap with respect to the benchmark - training separately on each test distribution. We refer to the training with data augmentation as a strong baseline that we seek to improve upon with our test-time adaptation method.

(IV) Domain Generalization Methods

Like data augmentation, we observed that all meta-learning based domain generalization approaches also substantially improved the performance over the baseline, for cases where the training and test distributions consisted of images acquired with the same imaging protocol, with differing scanners. However, a gap to the benchmark still remains. This shows the difficulty of learning distribution-invariant, task-performant models. We believe that this also vindicates our hypothesis that test-time adaptation is necessary for achieving good performance on test distributions that are unseen during training.

(V) Post-Processing Methods

We observed that post-processing with the DAE brought about substantial improvements over strong baseline, especially for the prostate datasets. However, the DAE post-processing lead to performance degradation on the brain datasets. We believe that this can be attributed to the fact that DAEs map their inputs to a plausible segmentation, however, that segmentation may not necessarily be tied to the test image. Furthermore, the post-processing method did not work when the source and target domains are acquired with different protocols or imaging modalities. In such cases, the pre-trained CNN predicted highly corrupted segmentations, which cannot be seen as samples from the DAE's training input distribution. Therefore, post-processing with the trained DAE could not improve the segmentation accuracy. This can be seen for the brain datasets, when the test distribution, HCP-T2, consists of T2w images, while the training distribution, HCP-T1, consists of T1w images.

(VI) Test-Time Adaptation

TTA provided substantial performance gains over competing methods across all datasets for brain and prostate anatomies. For the cardiac dataset, we observed that the strong baseline already provided a fairly good segmentation. The proposed method preserved this performance, but could not further improve it.

The improvement in Dice scores with the proposed method as compared to post-processing using the trained DAEs was statistically significant for 3 out of the 5 test datasets (marked with * in the Table), as measured using a paired Permutation test with 100000 permutations. For the other 2 test datasets, we obtained similar results upon direct post-processing as with the TTA using DAEs.

Qualitative results: Fig 5.3 also reveals similarly substantial improvements over strong baseline, especially for the brain and prostate datasets. It can be seen that TTA improved the predicted segmentation by, for instance, correcting predictions that are contextually misplaced, completing organ shapes and removing outliers.

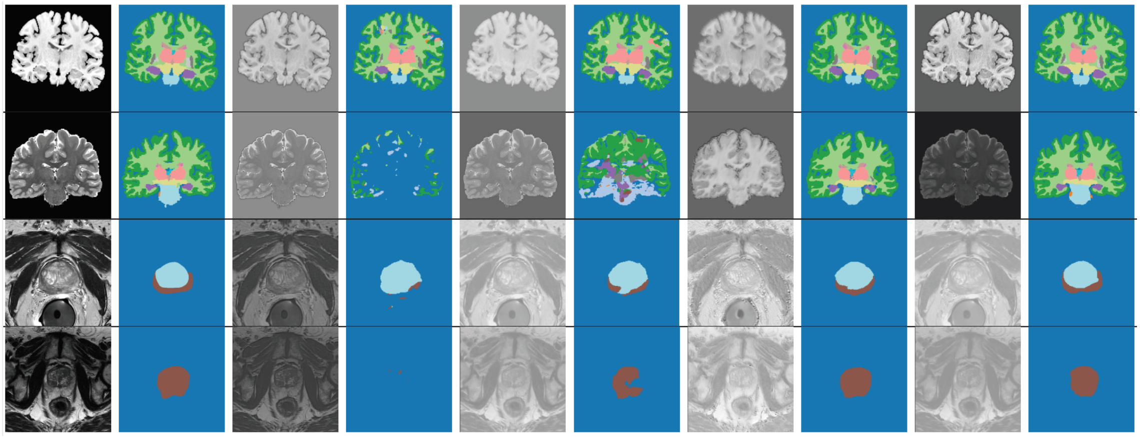


Figure 5.3: Qualitative results. Rows show results from test distributions for different anatomies. The first and second columns show test images and their ground truth segmentation respectively. After this, from left to right, normalized image and predicted segmentation pairs are shown for training with baseline (supervised learning of training distribution), strong baseline (supervised learning on training distribution with extensive data augmentation), proposed method (TTA using DAEs) and benchmark (supervised learning on test distribution).

(VII) Unsupervised Domain Adaptation Methods

Table 5.3 shows the results of our UDA experiments. Despite our persistent efforts, application of the method by [88]↑to large domain changes (modality change) and of the method by [90]↑to small domain changes (scanner changes within the same modality) led to poorer accuracies that the strong

DICE	Brain			Prostate (whole)			Prostate (sep.)		Heart	
Center / Method	HCP-T1	ABIDE-AC	HCP-T2	RUNMC	USZ	PROMISE12	RUNMC	USZ	ACDC	RVSC
Baseline and Benchmark										
Baseline	0.85	0.59	0.11	0.84	0.59	0.61	0.72	0.54	0.82	0.67
Benchmark	-	0.90	0.87	-	0.82	0.83	-	0.73	-	0.81
Strong Baseline - Data Augmentation										
Baseline + DA [101]†	0.87	0.75	0.08	0.91	0.77	0.79	0.82	0.66	0.83	0.74
Other Domain Generalization Methods										
MLDG [110]†	0.87	0.69	0.07	0.91	0.77	0.76	0.82	0.66	0.84	0.70
MASF [111]†	0.87	0.69	0.07	0.91	0.75	0.78	0.82	0.64	0.84	0.70
MLDGTS [112]†	0.88	0.73	0.07	0.91	0.71	0.76	0.82	0.61	0.83	0.36
Post-Processing Methods										
Strong Baseline + Post-Proc. [143]†	-	0.71	0.11	-	**0.79**	0.82	-	**0.68**	-	**0.75**
Proposed Method - TTA-DAE										
Strong Baseline + TTA	-	**0.80***	**0.73***	-	**0.79**	**0.86***	-	**0.68**	-	0.74

Hausdorff Distance	Brain			Prostate (whole)			Prostate (sep.)		Heart	
Center / Method	HCP-T1	ABIDE-AC	HCP-T2	RUNMC	USZ	PROMISE12	RUNMC	USZ	ACDC	RVSC
Baseline and Benchmark										
Baseline	9.1	34.3	52.1	16.7	147.9	69.0	19.1	146.3	15.9	16.9
Benchmark	-	1.4	2.3	-	21.9	14.3	-	22.4	-	3.5
Strong Baseline - Data Augmentation										
Baseline + DA [101]†	2.1	18.0	55.8	3.6	55.8	26.9	5.3	53.3	6.2	13.9
Other Domain Generalization Methods										
MLDG [110]†	2.5	21.2	52.3	2.5	43.5	32.3	4.7	42.8	7.7	22.5
MASF [111]†	2.1	18.5	57.9	5.6	90.2	45.8	6.6	87.3	6.2	15.9
MLDGTS [112]†	2.2	13.5	55.0	3.6	93.9	36.6	6.5	92.1	8.4	42.8
Post-Processing Methods										
Strong Baseline + Post-Proc. [143]†	-	13.2	53.6	-	38.9	14.3	-	38.5	-	**9.5**
Proposed Method - TTA-DAE										
Strong Baseline + TTA	-	**10.1**	**21.5**	-	**28.1**	**9.5**	-	**31.4**	-	15.3

Table 5.2: Quantitative Results: DICE in the top sub-table and 95^{th} percentile Hausdorff distance in the bottom sub-table. Mean results over all foreground labels, all test subjects and over 3 experiment runs are shown. For the prostate datasets, results are shown for whole gland segmentation (whole) as well as averaged over the central gland and peripheral zone (sep.). The rows show different training / adaptation strategies and are described in the text. For the proposed method, the * next to Dice scores denotes statistical significance over 'Strong Baseline + Post-Proc.' (Permutation test with a threshold value of 0.01).

baseline. To the best of our knowledge, this observation is consistent with the distribution shifts in the datasets presented in these works as well as other

Anatomy	Brain		Prostate (whole)		Heart
Center Method	ABIDE-AC	HCP-T2	USZ	PROMISE12	RVSC
Baseline + DA [101]$^\uparrow$(Strong baseline)	0.75	0.08	0.78	0.79	0.74
UDA - Invariant features [88]$^\uparrow$	0.80	0.08	0.79	0.80	0.75
UDA - Image-to-Image translation [90]$^\uparrow$	0.64	0.81	0.69	0.75	0.17
TTA (Proposed)	0.80	0.73	0.79	0.86	0.74
Benchmark	0.90	0.87	0.82	0.83	0.81

Anatomy	Brain		Prostate (whole)		Heart
Center Method	ABIDE-AC	HCP-T2	USZ	PROMISE12	RVSC
Baseline + DA [101]$^\uparrow$(Strong baseline)	0.75	0.08	0.77	0.75	0.72
UDA - Invariant features [88]$^\uparrow$	0.79	0.08	0.79	0.77	0.72
UDA - Image-to-Image translation [90]$^\uparrow$	0.65	0.82	0.61	0.77	0.25
TTA (Proposed)	-	-	-	-	-
Benchmark	-	-	-	-	-

Table 5.3: Dice scores for comparison of the proposed method with unsupervised domain adaptation (UDA) methods. The top sub-table shows results for the test dataset from the test distribution, $\mathcal{D}_{ts}^{UL,ts}$, while the bottom sub-table shows results for the training dataset from the test distribution, $\mathcal{D}_{ts}^{UL,tr}$. These results show that the proposed test-time adaptation method performs comparably with the best performing UDA methods, with the additional critical benefit of not requiring the labelled training distribution dataset while adapting for each new test distribution.

UDA works that follow similar ideas.

Remarkably, it can be seen that not using training distribution labelled dataset does not hinder the TTA method; it achieves comparable results to the best performing UDA methods, especially for the scanner change related DS. The image-to-image translation method using cycleGAN [90] provided better performance than TTA for the case with large DS (modality change), but lead to poorer results than the baseline for smaller DS (scanner changes).

Comparing the two sub-tables in Table 5.3, we see that within the UDA methods, the performance gains for $\mathcal{D}_{ts}^{UL,tr}$ are similar to that for $\mathcal{D}_{ts}^{UL,ts}$. Thus, using the same images for adaptation as well as testing did not lead to additional improvements in our experiments.

These results highlight that TTA can serve as a potent and more flexible alternative to UDA methods, for both small and large DS.

(VIII) Ablation Studies

The results of our ablation studies are shown in Table 5.4, and described be-

Anatomy	Brain			Prostate (whole)			Prostate (sep.)		Heart	
Method \ Center	HCP-T1	ABIDE-AC	HCP-T2	RUNMC	USZ	PROMISE12	RUNMC	USZ	ACDC	RVSC
Proposed Method - TTA-DAE										
Adapt ϕ, using DAE	-	0.80	0.73	-	0.79	0.86	-	0.68	-	0.74
Fast version of the proposed method										
Adapt ϕ, using DAE - Fast	-	0.80	0.73	-	0.79	0.84	-	0.68	-	0.75
Adapting all parameters for TTA										
Adapt ϕ, θ, using DAE	-	0.67	0.65	-	0.72	0.61	-	0.58	-	0.71
TTA using Ground Truth Labels										
Adapt ϕ, using GT labels	-	0.83	0.84	-	0.84	-	-	0.77	-	-
Iterative Post-Processing with DAE										
10 passes through DAE	-	0.63	0.11	-	0.79	0.83	-	0.69	-	0.73
100 passes through DAE	-	0.53	0.10	-	0.79	0.82	-	0.67	-	0.69

Anatomy	Brain			Prostate (whole)			Prostate (sep.)		Heart	
Method \ Center	HCP-T1	ABIDE-AC	HCP-T2	RUNMC	USZ	PROMISE12	RUNMC	USZ	ACDC	RVSC
Proposed Method - TTA-DAE										
Adapt ϕ, using DAE	-	10.1	21.5	-	28.1	9.5	-	31.4	-	15.3
Fast version of the proposed method										
Adapt ϕ, using DAE - Fast	-	9.2	19.2	-	30.7	9.8	-	33.9	-	12.7
Adapting all parameters for TTA										
Adapt ϕ, θ, using DAE	-	5.0	13.7	-	15.5	16.1	-	22.2	-	4.8
TTA using Ground Truth Labels										
Adapt ϕ, using GT labels	-	6.1	3.9	-	35.2	-	-	35.4	-	-
Iterative Post-Processing with DAE										
10 passes through DAE	-	15.1	56.5	-	28.5	11.6	-	30.8	-	7.2
100 passes through DAE	-	21.6	58.3	-	25.7	10.4	-	29.3	-	9.2

Table 5.4: Ablation Study: Upper sub-table shows DICE scores, and the lower one shows 95th percentile Hausdorff distance.

low.

1. **'Fast' TTA**: It can be seen that the TTA performance is immune to the change in the initialization strategy. This demonstrates that the proposed adaptation works in a stable manner for multiple test images, and is not dependent on a 'good' first test image from a test distribution. In practice, the fast version is more attractive due to the time saving.

2. **Effect of adapting all parameters $\{\theta, \phi\}$ for TTA**: It can be seen that this lead to a drop in segmentation accuracy in terms of Dice score, but improved the Hausdorff distance (in all test distributions, except PROMISE12 for prostate).

Qualitatively, we observed that the Dice scores deteriorated because the segmentations while becoming more plausible, became inaccurate around the edges. The Hausdorff distance, on the other hand, improved because the added flexibility allowed outliers to be removed more effectively. Overall, we believe that accurate segmentations around organ edges are more valuable than removing extreme outliers (which can be removed by other post-processing steps, if required). Thus, we believe that this experiment showcases the importance of freeze a majority of the parameters at the values obtained from the initial supervised learning.

3. **Expressiveness of N_ϕ**: Naturally, using the ground truth labels for guiding TTA led to improvements in accuracy in all cases, as compared to the strong baseline. However, it is interesting to note that for the brain test distribution datasets, the resulting accuracies were inferior to the benchmark (that is, training specialized CNNs separately for each test distribution). This shows that despite TTA, some bias towards the training distribution may remain in the normalized-image-to-segmentation CNN, S_θ.

4. **How does TTA compare to iterative post-processing?**: The results of this experiment are shown in last two rows of Table 5.4 and in Fig. 5.4, for different number of passes through the DAE. We observed that such a post-processing approach could not improve segmentation accuracy as much as TTA. On the contrary, for the brain datasets, where the segmentations are more complicated than other anatomies due to the presence of multiple structures, the post-processing with multiple DAE passes worsened the segmentation accuracy. We believe that this might be because the DAE output, although generally plausible according the labels in the training distribution dataset, is not necessarily tied to the input image in question. The proposed method constrains the predicted segmentations to be tied to the input image by: (1) freezing the deep sub-network, S_θ and (2) keeping the adaptable sub-network, N_ϕ, relatively shallow. Also, the limited flexibility for the adaptation guards against potential errors of the DAE such as the one seen fourth column for the brain dataset in Fig. 5.4.

(IX) Convergence of TTA

We find that the adaptation converges across the more than 100 test volumetric images across different anatomies and test distributions, and across multiple runs of the experiments. Fig. 5.6 shows that TTA convergences reliably for different test distributions.

Fig. 5.5 shows an example evolution during TTA iterations of (i) the normalized images, (ii) the predicted segmentations and (iii) the denoised segmen-

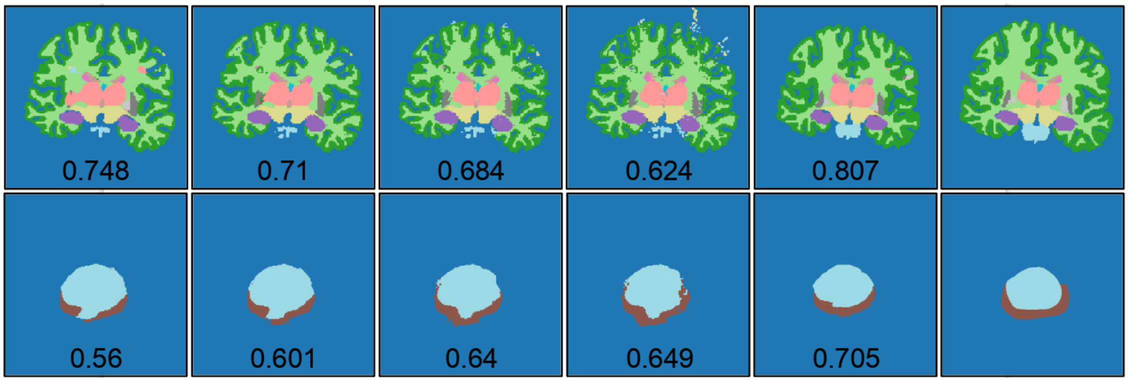


Figure 5.4: Multiple passes through the DAE do not suffice to improve segmentation accuracy. From left to right: Initial prediction by the strong baseline, followed by 1, 10 and 100 passes through the DAE, followed by the TTA prediction and finally, the ground truth. Top and bottom rows show results for the ABIDE-AC and USZ datasets, respectively, and the numbers below the segmentations are the corresponding volumetric averaged foreground Dice scores.

tations suggested by the DAE for improving segmentation performance.

Fig.5.7 shows the correlation between (a) the Dice between the predicted and ground truth segmentation and (b) the Dice between the DAE input and DAE outputs. It can been seen that these two Dice scores are correlated, thus justifying the choice of using the latter as the TTA model selection criterion (see last paragraph of Sec. 5.3.3).

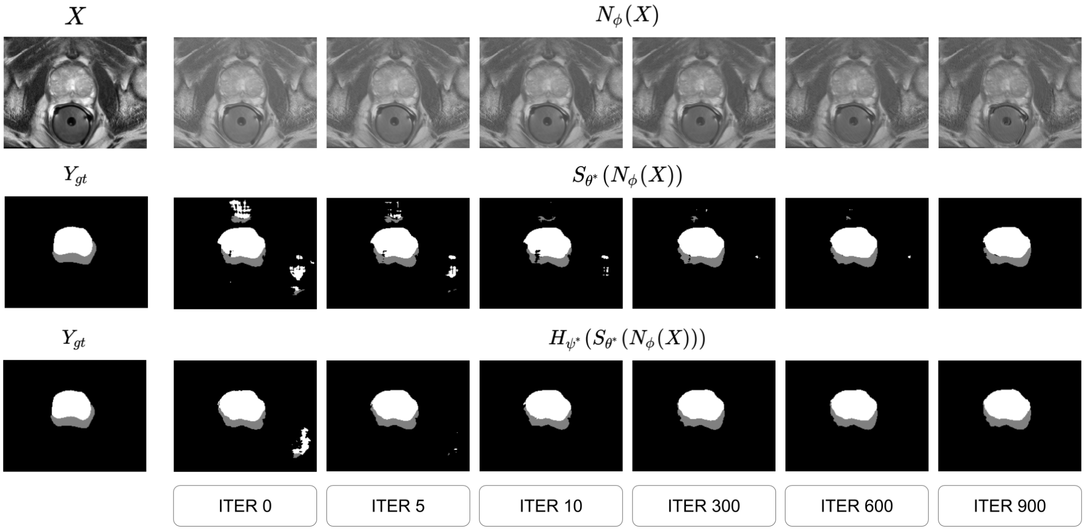


Figure 5.5: Evolution during TTA iterations of (i) the normalized images (top), (ii) the predicted segmentations (middle) and (iii) the denoised segmentations suggested by the DAE for improving segmentation performance (bottom).

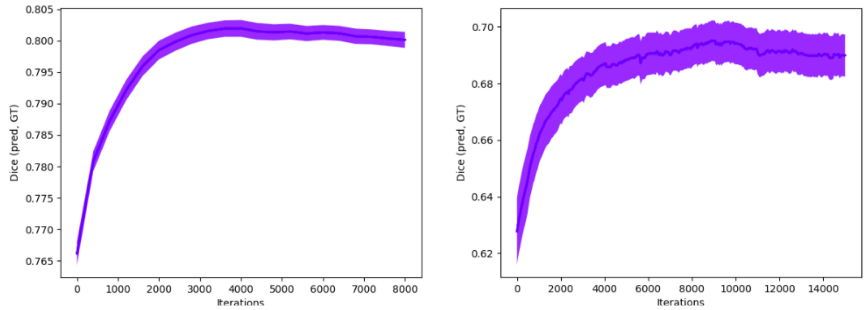


Figure 5.6: Convergence of TTA for different anatomies and test distributions: ABIDE-AC (left), USZ (right). Mean and 0.1 * std. deviation over different test subjects, of the Dice between the predicted and ground truth segmentation, as a function of TTA iterations.

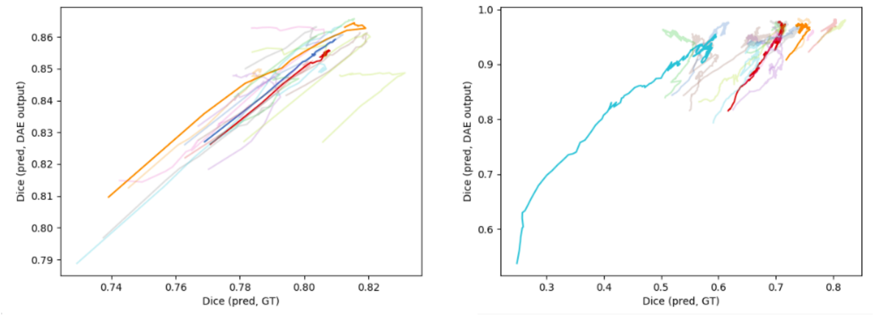


Figure 5.7: Correlation between (a) the Dice between the predicted and ground truth segmentation and (b) the Dice between the DAE input and DAE outputs. Each color represents TTA of a single test subject. For clarity, a few subjects are shown in opaque colours, while others are faded. ABIDE-AC (left), USZ (right).

5.5 Discussion

In this chapter, we proposed a test-time adaptation method for improving robustness to acquisition-related DS in medical image segmentation. The method consists of two main ideas. Firstly, we introduce an adaptable per-image normalization module into a segmentation CNN. We believe that such per-image adaptability may be crucial for developing robust analysis tools that can be deployed in the clinic. Secondly, the proposed TTA is driven by using denoising autoencoders, that incentivize plausible segmentation predictions. Experiments with multiple datasets and anatomies demonstrate the promise and generality of the method, over other approaches such as data augmentation, meta-learning and unsupervised domain adaptation. The proposed method yields promising improvements while segmenting images from completely unseen scanners and / or protocols.

In the following paragraphs, we discuss some avenues that could be potentially interesting for further improving TTA performance, and for ultimately closing the gap to the benchmark - i.e. training a separate CNN for each test distribution.

Noising strategy for DAE training

One of the main assumptions of our work is that the incorrect segmentations predicted by a CNN for a test image from an unseen test distribution can be considered to be from the training input distribution of the DAE. The validity of this assumption is crucial for the output of the DAE to be reliable. In this work, we chose a heuristic strategy for corrupting segmentation labels that the DAE seeks to denoise. Thus, if the corrupted prediction for a certain test image is not represented by the chosen heuristic strategy, the corresponding DAE outputs may not be suitable for TTA for that test image.[4] The DAE performance can potentially be further improved if a strategy can be devised to obtain noisy labels from the trained segmentation CNN. A potential way of doing this might be to train the segmentation CNN on the training distribution without data augmentation, and then to use the predictions of this CNN on intensity transformed training distribution images (for instance, via gamma transformations) as noisy segmentations.

Capturing segmentation denoising uncertainty

A certain predicted corrupted segmentation could by generated from multiple plausible clean segmentations. A common assumption in the DAE literature [221]↑is that a trained DAE provides samples from the posterior distribution over clean segmentations Y, given a corrupted segmentation Y_c, $P(Y|Y_c)$. However, this assumption may not necessarily hold as the DAE training procedure does not ensure this. Instead, the DAE is trained to output only one clean segmentation rather than several possible denoised segmentations. The performance of the DAE, and thus of the TTA, can be potentially further improved by training the DAE to obtain such behaviour [222]↑.

Modeling the distribution of normalized images

We trained the overall segmentation CNN, $S_\theta(N_\phi(.))$, with extensive data augmentation and then adapted the normalization sub-network, $N_\phi(.)$, for each test image. We emphasize that the output of $N_\phi(.)$ is an intermediate representation Z, different from the input images X, even for images from the training distribution(s). Thus, the proposed TTA is not expected to tune $N_\phi(.)$ such that it acts as a translator from the test to the training distribution. In-

---

[4] This concern is one of the main motivations for the alternative TTA method proposed in the next chapter.

stead, we expect TTA to tune $N_\phi(.)$ such that is maps the given test image to the distribution of normalized versions of the training distribution images.

This observation suggests an alternative strategy to drive TTA - by modelling the distribution of the normalized training images, followed by asking the test image to have a normalized representation that adheres the learned distribution. In a nutshell, this is the TTA strategy followed in the method described in the next chapter.

Furthermore, we note that it may be possible to impose additional constraints such that the normalized representation has desirable properties. For instance, we observed in our analysis experiments that carrying out TTA even with ground truth labels from the test distribution provides sub-optimal results as compared to the benchmark. This indicates the presence of distribution-specific signals in the space of normalized images. To overcome this problem, domain generalization strategies may be employed to achieve distribution invariance in this space during the initial supervised learning on the training distributions. Alternatively, specifically for MR images, it may be interesting to incorporate methods for quantitative mapping of MR tissue parameters [223]↑to achieve such acquisition-independent representations in the space of normalized images.

Affine registration for large domain shifts

The proposed atlas-based initialization for large DS was only evaluated with brain images. Additionally, as both the training distribution (T1w) and test distribution (T2w) images used from the HCP dataset were already rigidly aligned, the affine registration step for atlas creation could be skipped. Such affine registration would be required for other test distributions of brain images, as well as while applying the method to large DS in other anatomies. The proposed method's reliance on alignment with only linear transformations (rather that deformable registration) might facilitate applications in other anatomies as well, but we leave this evaluation to future work.

Test-time adaptation in a Bayesian framework

A possible extension of our method might be to consider TTA of a supervised CNN in the Bayesian framework, often used in unsupervised learning methods [30]↑. This would entail the use of an explicit prior model in the space of segmentation labels, $P(Z)$, as well as a likelihood model, $P(X|Z)$, with TTA being driven with the aim of maximizing the resulting posterior, $P(Z|X)$.

Time required for test-time adaptation

The per-image flexibility offered by our method comes at the cost of the additional time required for such adaptation. After the first image of a particular scanner / protocol, the adaptation requires about 12 minutes for each 3D image, with our experimental implementation. Despite potential for improvement in terms of time efficiency, the proposed TTA does introduce an additional optimization routine for each test image and thereby compromises on the fast-inference advantage of CNNs. Nonetheless, we believe that such a time requirement is relatively modest and reasonable for general usage in clinical practice.

Chapter 6

A Task-Agnostic Test-Time Adaptation Approach for Robust Medical Image Analysis

This chapter is based on the publication "A field of experts prior for adapting neural networks at test time" [224]↑. In this work, we extend the TTA approach of Chapter 5 such that it can be generically applied to multiple tasks.

6.1 Introduction

Test-Time Adaptation (TTA) Recap

In TTA, the parameters of a previously trained CNN are adapted for each test image. The subset of the parameters that get adapted per test image is a design choice. Noting that acquisition-related DS manifest as contrast variations, one approach is to design the CNN as a concatenation of a shallow, image-specific contrast normalization CNN, $z = N_\phi(x)$, followed by a deep task CNN that is shared by all training and test images, $y = S_\theta(z)$. Here, x is the input image, z is the normalized image, and y is the output (e.g. segmentation, deformation field, enhanced image).

The image-specific parameters, ϕ, are adapted by requiring adherence to a prior model, H_ψ, either in the output space [67]↑or in the feature space [68]↑. H_ψ encourages similarity between outputs or features of the test image with those of the training images. It is itself modelled using a CNN and trained in a self-supervised manner - as a denoising autoencoder (DAE) in [67]↑and as an autoencoder (AE) in [68]↑.

The DS problem in H_ψ

In this work, we scrutinize the prior model, H_ψ, which is a key component in

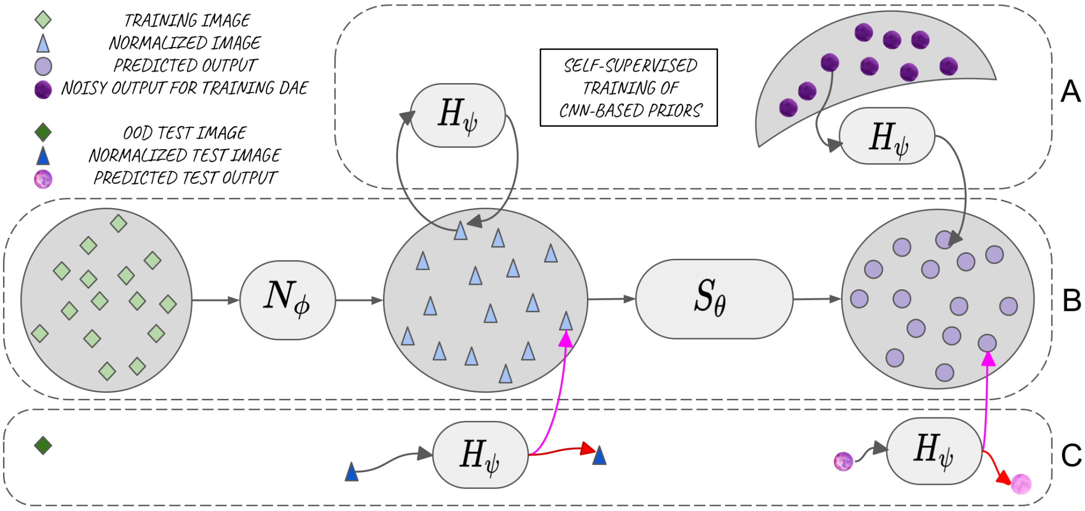


Figure 6.1: An illustrative schematic of the DS problem in CNN-based helper models. The figure is divided into 3 horizontal slabs (enclosed with dashed boundaries). Slab B shows the mapping of the inputs (green) to the outputs (purple), via the normalized features (blue). Slab A shows the training of prior models (autoencoder (center) [68]↑, denoising autoencoder (right) [67]↑) to be used for TTA: the AE is trained to auto-encode features of training images and the DAE is trained to denoise corrupted outputs (from a specific corruption distribution indicated by the crescent). Finally, slab C shows the desirable behaviour (pink arrows) and potential failure cases (red arrows) when the trained prior models are used to guide TTA.

tackling the DS problem via TTA. Consider what happens when TTA is used to improve a CNN's prediction accuracy in the presence of acquisition-related DS. At the beginning of TTA iterations, the test features (outputs) are likely to be dissimilar to the features (outputs) corresponding to the training images. Indeed, this is symptomatic of the CNN's poor performance on OOD images. The main assumption of TTA methods like [67]↑, [68]↑is that H_ψ is capable of mapping such features (outputs) to ones that are similar to features (outputs) observed during training.

We argue that, if H_ψ is modelled with a CNN, it is likely to be vulnerable to a DS problem of its own - that is, the outputs of H_ψ may be unreliable when its test inputs are from a different distribution as compared to its training inputs. An illustrative schematic of this problem is shown in Fig. 6.1. AEs in [68]↑, which are trained to auto-encode features of training images, are not guaranteed to transform the features of test images to be like the features of training images. Similarly, DAEs in [67]↑, trained to denoise corrupted outputs corresponding to a particular corruption distribution, may be unable to denoise outputs with different corruption patterns.

Lack of DS Robustness in Density Estimation Models

Although DAEs (for arbitrary corruption distributions) and AEs lack a strict probabilistic underpinning, the aforementioned TTA approaches can be roughly thought of as learning a probabilistic model of the training features (outputs), and then increasing the likelihood of the test features (outputs) under the trained model. We argue that even if CNN-based unsupervised density estimation models are used as the prior, they too are likely to suffer from the DS problem [164]↑, [225]↑. For instance, one approach for TTA might be to train variational autoencoders (VAEs) to model the distribution of features of the training images, and to modify the test image's features such that their likelihood under the trained VAE increases. VAEs may even assign higher likelihood values to OOD samples than samples from their training distribution [164]↑. Such behaviour may render them unsuitable for TTA.

Overview of the proposed method

In this work, we propose two main changes as compared to recent TTA works.

(1) Distribution Matching for TTA: Instead of driving TTA by minimizing the reconstruction loss of the prior model, H_ψ, we propose to match the distribution of 2D slices of a volumetric test image with the distribution of slices of training images. The distribution matching is done in the space of the normalized images, z.

(2) A Field-of-Experts Prior: Noting the lack of DS robustness in CNN-based prior models for driving TTA, we posit that simpler prior models may (a) suffice to improve task performance under the considered acquisition-related DS, while (b) themselves being more robust to DS as compared to CNN-based priors. With this motivation, we model the distribution of the normalized training images, z, using a Field of Experts (FoEs) [226]↑formulation. FoEs (described in more detail in Sec. 6.2) combine ideas of Markov random fields (MRFs) [227]↑and Product of Experts (PoEs) [228]↑. FoEs enable modeling of complex distributions as a product of several simpler distributions. The simple distributions are those of the outputs of so-called expert functions, which are typically formulated as scalar functions of image patches. We propose to use the task-specific filters learned in S_θ as the FoE experts (see Sec. 6.3.3 for details). Further, we augment the FoE model with additional experts - projections onto principal components of patches in the last layer of S_θ (Sec. 6.3.5).

For TTA, we adapt the normalization module N_ϕ, so as to match the individual expert distributions of the test and training images, for all experts in the FoE model.

Summary of contributions To summarize, we consider the acquisition-related

DS problem in CNN-based medical image analyses and make the following contributions in this work: (1) we propose distribution matching for TTA, (2) we model the distribution of normalized images, z, using a FoE model, with the task-specific CNN filters acting as the expert functions, and (3) we augment the FoE model with PCA-based expert functions.

We support these technical contributions with an extensive validation on 5 image segmentation tasks, using data from 17 centers, and an image registration task, using data from 3 centers. To the best of our knowledge, this is the first work in the literature that evaluates the TTA setting on such a large variety of anatomies and tasks for medical image analysis. The results of these experiments help us organize the current TTA literature, including the proposed method, along three axes. (1) Applicability to multiple tasks: some of the existing TTA methods are task-dependent. The proposed method relieves this constraint, and provides a general approach that can used in multiple tasks. As compared to existing task-agnostic methods, the proposed method provides similar performance for image registration and superior performance for image segmentation. (2) Performance in segmentation of anomalies: we find that DS robustness issue is particularly difficult for lesion datasets. Here, all of the existing TTA methods either fail to improve performance, and several methods even lead to performance degradation as compared to the baseline. The proposed method provides substantial performance gains in this challenging scenario. (3) Performance in segmentation of healthy tissues: in this scenario, our experiments indicate that methods specifically designed for handling distribution shifts in image segmentation outperform more general TTA methods, including the proposed method.

6.2 Background

6.2.1 Markov Random Fields (MRFs)

MRFs [227]↑express a probability density function of an image, z, as an energy-based model:

$$p(z) = \frac{1}{\mathcal{C}} \exp(-E(z)) \tag{6.1}$$

where $\mathcal{C}$ is a normalization constant. The energy of the image is defined as the sum of energies (potential functions) of all constituent $\mathcal{R}^{k \times k}$ patches (cliques), z_k:

$$E(z) = \sum_{\mathcal{K}} E(z_k) \tag{6.2}$$

where $\mathcal{K}$ denotes the set of all $k \times k$ patches. Typically, the energy function $E(z_k)$ is defined over relatively small patches and is hand-crafted - for instance, to encode smoothness.

6.2.2 Field of Experts (FoEs)

FoEs [226]↑extend the MRF idea by learning the energy function from data. Specifically, the energy of image patches, z_k, is written in the Product-of-Experts (PoE) framework [228]↑, [229]↑:

$$E(z_k) = -\sum_{j=1}^{J} \log p(f_j(z_k); \alpha_j) \tag{6.3}$$

Substituting this into the MRF expressions shown in Sec. 6.2.1, the energy of the total image, z, becomes

$$E(z) = -\sum_{\mathcal{K}} \sum_{j=1}^{J} \log p(f_j(z_k); \alpha_j) \tag{6.4}$$

The corresponding probability density function of the image, z, becomes

$$p(z) = \frac{1}{\mathcal{C}} \prod_{k \in \mathcal{K}} \prod_{j=1}^{J} p(f_j(z_k); \alpha_j) \tag{6.5}$$

Here, $f_j : \mathcal{R}^{k \times k} \rightarrow \mathcal{R}$ are expert functions, and α_j are parameters of the 1D distributions of experts' scalar outputs. The key idea in PoE and thus, FoE models is that each expert models a particular low-dimensional aspect of the high-dimensional data. Due to the product formulation, only data points that are assigned high probability by all experts are likely under the model. In [228]↑, [229]↑, [226]↑, f_j and α_j are learned using an algorithm known as contrastive divergence, such that images in a training dataset are assigned low energy values, and all other points in the image space are assigned high energy values.

6.3 Method

6.3.1 Splitting Parameters into Image-Agnostic and Image-Specific

We follow the parameter splitting strategy that is proposed in the previous method, TTA-DAE (chapter 5). Please refer to Sec. 5.3.1 for a detailed reasoning behind this choice of parameter splitting. We provide a brief summary here.

An image, x, is passed through a shallow normalization module, N_ϕ, which outputs a normalized image, z. N_ϕ consists of a few (2-4) convolutional layers with relatively small kernel size (1-3) and stride 1, and outputs z, which is a feature with the same spatial dimensionality and the same number of channels as x. z is passed through a deep CNN, S_θ, which produces the output y. y is formulated as per the task at hand - for instance, it can be a segmentation mask, a deformation field, a super-resolved image, etc. Fig. 6.2 shows a representative CNN architecture in this framework. We consider 2D CNNs, but in principle, the method may be extended to 3D architectures as well.

6.3.2 Supervised Learning on Training Distribution

Similar to TTA-DAE (Sec. 5.3.2), we train S_θ and N_ϕ using labelled input-output pairs from the training distribution. At test-time, S_θ is fixed, while N_ϕ is adapted for each test volumetric image.

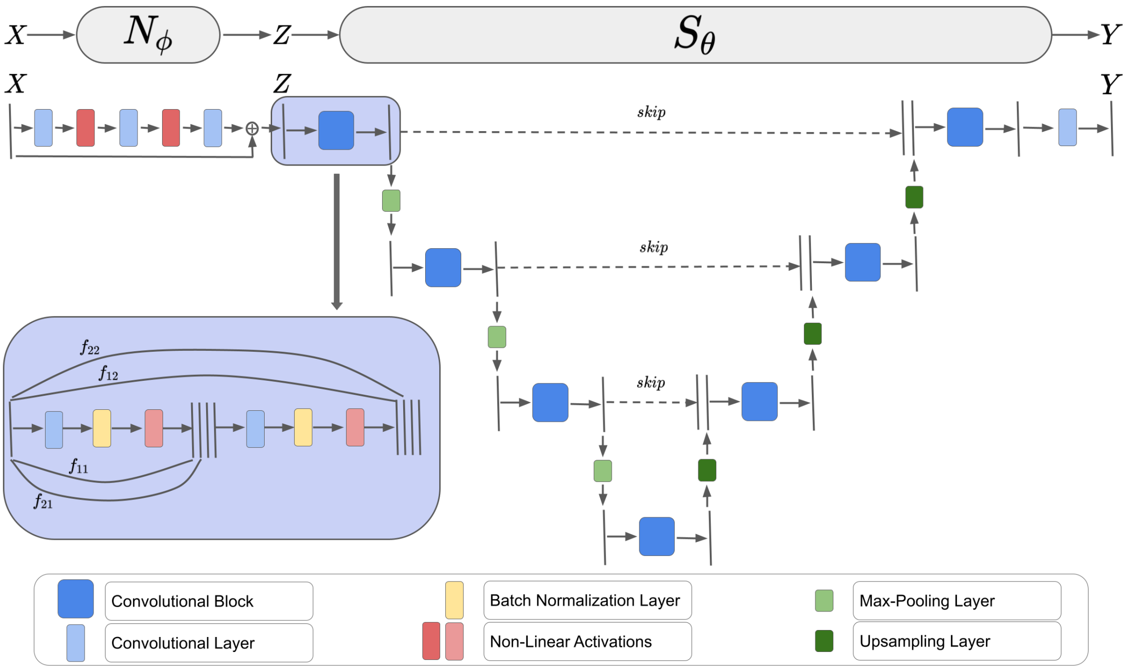


Figure 6.2: Representative schematic of a test-time adaptable CNN. We follow the same parameter splitting strategy as in TTA-DAE: A shallow normalization sub-network, N_ϕ, is image-specific and is adapted for each test image, while a deep segmentation sub-network, S_θ segments the normalized image Z and is shared by all training and test images. Each convolutional block in S_θ has the same architecture as the networks in Fig. 5.1 and Fig. 4.1, but is shown in more detail here. Specifically, the notation f_{cl} is shown, which represents the function taking as input the normalized image Z, and outputting the c^{th} channel of the l^{th} layer of S_θ.

6.3.3 FoE-CNN: A New Helper Model for TTA

In this work, we drive TTA by matching the distribution in the space of normalized images, z, of the given volumetric test image with that of the training images. In this section, we describe our distribution modeling approach. In Sec. 6.3.4, we describe how to match the modelled distributions.

Notation

Consider a representative architecture for S_θ shown in Fig. 6.2. The first convolutional block has been highlighted in the figure to exemplify the following notation: We use f_l: $\mathcal{R}^{N_x \times N_y} \rightarrow \mathcal{R}^{N_{xl} \times N_{yl} \times C_l}$ to denote the function that takes as input the normalized image z, and outputs the features of the l^{th} convolutional layer of S_θ. Further, we use f_{cl}: $\mathcal{R}^{N_x \times N_y} \rightarrow \mathcal{R}^{N_{xl} \times N_{yl}}$ to denote the function that takes as input z, and outputs the c^{th} channel of the l^{th} convolutional layer of S_θ.

If k_l is the receptive field at f_{cl} with respect to z, each pixel in the output of f_{cl} can be seen as a 1D projection of a $k_l \times k_l$ patch of z, i.e., an expert function.

For ease of reading, we overload the notation f_{cl} to indicate two things: (a) if written as $f_{cl}(z)$, it indicates the function that takes as input z, and outputs the c^{th} channel of the l^{th} convolutional layer of S_θ and (b) if written as $f_{cl}(z_{k_l})$, it indicates the function that takes as input an image patch of z of size $k_l \times k_l$, and outputs the corresponding pixel in the c^{th} channel of the l^{th} convolutional layer of S_θ.

Distribution Modeling

We model the distribution of normalized images, z, using the FoE formulation (Sec. 6.2.2) with the 3 modifications.

(i) **Multiple patch sizes**: Firstly, note that in the original FoE model, the energy function is defined in terms of input patches of a single patch size. We consider multiple patch sizes to define the energy. Specifically, if S_θ consists of L convolutional layers, we consider L patch sizes - namely, the receptive fields of all the convolutional layers of S_θ.

$$E(z) = \sum_{l=1}^{L} \sum_{\mathcal{K}_l} E((z_{k_l})) \tag{6.6}$$

where $\mathcal{K}_l$ denotes the set of all $k_l \times k_l$ patches.

(ii) **Task-specific experts**: Secondly, we define the energy function for each patch size, using a separate PoE model. However, unlike [226]↑, we do not learn the expert functions using contrastive divergence. Instead, we construct a task-specific FoE model by using the functions f_{cl} of S_θ as C_l experts to describe the energy of patches of z of size $k_l * k_l$:

$$E(z_{k_l}) = -\sum_{c=1}^{C_l} \log\ p(f_{cl}(z_{k_l}; \alpha_{cl})) \tag{6.7}$$

As previously noted, $f_{cl}(z_{k_l})$ are individual pixels of the c^{th} channel of the l^{th} convolutional layer of S_θ. Thus, $p(f_{cl}(z_{k_l}; \alpha_{cl}))$ is the 1D distribution of these pixel values, and α_{cl} are its parameters. Combining Eqns 6.7 and 6.6, and inserting the resulting energy function into the FoE formulation (Sec. 6.2.2), the corresponding PDF of the normalized images can be written as:

$$p(z) = \frac{1}{\mathcal{C}} \prod_{l=1}^{L} \prod_{c=1}^{C_l} \prod_{\mathcal{K}_l} p(f_{cl}(z_{k_l})) \tag{6.8}$$

Change of notation: For ease of reading, let us denote expert outputs, $f_{cl}(z_{k_l})$, by u and their distribution, $p(f_{cl}(z_{k_l}); \alpha_{cl})$, by $p_{cl}(u; \alpha_{cl})$. Also, note that the product over $k_l * k_l$ patches of Z is the product over the pixels of f_{cl}. Thus, we have:

$$p(z) = \frac{1}{\mathcal{C}} \prod_{l=1}^{L} \prod_{c=1}^{C_l} \prod_{i=1}^{N_{xl} * N_{yl}} p_{cl}(u_i; \alpha_{cl}) \tag{6.9}$$

The functions learned in S_θ act as task-specific experts. We hypothesize that matching the distributions of the outputs of such experts during TTA is likely to be beneficial for improving the task performance for the test images.

(iii) **Estimation of experts' distributions**: We approximate the expert distributions, $p_{cl}(u; \alpha_{cl})$, as 1D Gaussian distributions, with $\alpha_{cl} = \{\mu_{cl}, \sigma_{cl}\}$:

$$\begin{aligned} &p_{cl}(u; \alpha_{cl}) = \mathcal{N}(\mu_{cl}, \sigma_{cl}) \\ &\mu_{cl} = \tfrac{1}{N_z} \textstyle\sum_z \tfrac{1}{N_{xl} * N_{yl}} \sum_i u_i,\ \sigma_{cl}^2 = \tfrac{1}{N_z} \sum_z \tfrac{1}{N_{xl} * N_{yl}} \sum_i (u_i - \mu_{cl})^2 \end{aligned} \tag{6.10}$$

Here, the outer sum, $\sum_z$, is over all samples of normalized images z, and the inner sum, $\sum_i$, is over all pixels of the feature at the c^{th} channel of the l^{th} layer.

Eqn. 6.9 defines the complete field of CNN experts probability model (FoE-CNN) of the normalized images, z, with the individual expert PDFs given either by Eqn. 6.10.

In Sec. 6.4.4 (analysis experiments), we analyze the effect on TTA of modelling $p_{cl}(u)$ using kernel density estimation (KDE). While this approach can capture higher-order moments of the distributions, we observed that the resulting PDFs were relatively similar to their Gaussian approximations. Thus, for simplicity, we propose to use the Gaussian approximation in the method, and show the effect of using KDE in the appendix.

6.3.4 How to Drive TTA using the FoE-CNN Model?

We propose to use to the FoE-CNN model for TTA in the following setting: at the training site (e.g. hospital), multiple labelled volumetric images are available from the training distribution, but at the test site, we would like to adapt the model for each volumetric test image separately. Therefore, we consider subject-specific distributions $p^s(z)$ (Eqn. 6.9), consisting of subject-specific 1D PDFs, $p^s_{cl}(u)$ (Eqn. 6.10). That is, after training N_ϕ and S_θ using data from the training distribution, we compute and save the 1D PDFs, $p^s_{cl}(u)$, for all channels of all layers, for all training subjects. These are transferred to the test site. A practical advantage here is that only summary statistics of the 1D distributions are transferred - this provides benefits in terms of privacy and memory requirements, as compared to transferring large CNN models or the training distribution images themselves. Now, for TTA, we have to make the following two design choices.

(i) **Log-likelihood maximization v/s Distribution matching**: Given a test subject t, there are two possible ways to carry out TTA. One option is to maximize the log-likelihood of the normalized image corresponding to the test image, under the FoE-CNN model computed for the training images. Further, since the distribution of the training subjects are also modelled subject-wise, we additionally take an expectation over the training subjects:

$$max_\phi \; E_{p(s)} \; [E_{p^t(z)} \; \log \; p^s(z)] \rightarrow max_\phi \; E_{p(s)} \; [E_{p^t(z)} \; \textstyle\sum_{l=1}^{L} \; \sum_{c=1}^{C_l} \; \sum_{i=1}^{N_{xl}*N_{yl}} \; \log \; p^s_{cl}(u_i)] \tag{6.11}$$

We approximate the expectation with respect to $p^t(z)$ using randomly chosen 2D slices of the test subject's volumetric image. A potential problem with this TTA formulation may be that it attract all pixels u_i towards the modes of p^s_{cl}. To circumvent this issue, we propose to model the distribution of the normalized images corresponding to the 2D slices of the test subject, $p^t(z)$, also using the FoE-CNN model (Eqn. 6.9). Now, a suitable divergence measure, D, between this and the distributions of the training subjects can be minimized.

$$min_\phi \; E_{p(s)} \; D(p^s(z), p^t(z)) \tag{6.12}$$

However, the normalization constant $\mathcal{C}$ in Eqn. 6.9 is intractable to compute and may be different for the two distributions. As well, commonly used divergence measures (such as f-divergences) require integration over the entire space over which the distributions are defined. Clearly, this is not possible for the high-dimensional normalized images, z. Therefore, for TTA, we match all the 1D expert distributions, p_{cl}, for all channels of all layers. That is, we minimize $L_{FoE-CNN}$ with respect to ϕ, where

$$L_{FoE-CNN} = E_{p(s)}\Big[\tfrac{1}{L} \textstyle\sum_{l=1}^{L} \tfrac{1}{C_l} \sum_{c=1}^{C_l} D(p^s_{cl}(u), p^t_{cl}(u)) \Big] \tag{6.13}$$

In particular, we minimize the KL-divergence between the individual 1D distributions for the training and test images. As the 1D PDFs are approximated as Gaussians, the KL-divergence can be computed in closed form. Further, for this choice of divergence measure, we show in Sec. 6.7.1 that minimizing the objective in Eqn. 6.12 is equal to minimizing the one in Eqn. 6.13 plus log of the normalization constant $\mathcal{C}$ for the test image.

(ii) **Incorporating information from multiple training subjects**: As mentioned previously, we consider subject-specific distributions of the normalized images. This provides us with two options for carrying out distribution matching for TTA: (a) minimize the divergence of the test subject's distribution with the expected distribution over all training subjects: $min\ D(E_{p(s)}[p^s_{cl}(u)], p^t_{cl}(u))$. (b) minimize the expected divergence of the test subject's distribution with the distribution of each training subject: $min\ E_{p(s)}[D(p^s_{cl}(u), p^t_{cl}(u))]$. For KL-divergence, we show in Appendix 6.7.2 that two objectives are related as follows:

$$\begin{aligned} &D_{KL}\big(E_{p(s)}[p^s_{cl}(u)], p^t_{cl}(u)\big) = \\ &- E_{p(s)}[D_{KL}(p^s_{cl}(u), E_{p(s)}[p^s_{cl}(u)])] + E_{p(s)}[D_{KL}(p^s_{cl}(u), p^t_{cl}(u))] \end{aligned} \tag{6.14}$$

As the first term on the right-hand side of Eqn 6.14 does not depend on the test image, TTA should, in principle, be equivalent for both ways of incorporating information from multiple training subjects. However, computing (b) in practice requires only one monte-carlo (MC) approximation, while computing (a) requires three MC approximations over the training subjects. Thus, the variance of (b) will be less than that of (a) [230]↑. With this reasoning, we choose (b) over (a) in the proposed TTA objective (Eqn 6.13).

The proposed TTA approach presents a general distribution matching framework, of which TTA methods that use test image(s) statistics in the BN layers of a trained CNN [131]↑, [132]↑are specific instances. While the replacement of batch normalization statistics is heuristically motivated in [131]↑, [132]↑, the 1D feature distribution matching in the proposed TTA strategy emerges from a principled framework. Further, as discussed in the following sub-section, the proposed framework naturally extends to include 1D distribution matching in the space of PCA loadings.

6.3.5 FoE-CNN-PCA: An Extended Helper Model

We note that the task-specific experts, f_{cl}, in the proposed probability model (Eqn. 6.9) take as inputs patches of increasing patch sizes. The experts f_{cL} have the largest receptive field, k_L, - thus, they model spatial correlations in $k_L \times k_L$ patches. Depending on the architecture of S_θ, this may or may not cover the entire spatial dimensionality of the normalized image z. We hypothesize that considering spatial correlations in even larger image patches may further improve the proposed TTA. Furthermore, even within the already considered patch sizes, the task-specific experts derived from S_θ may not necessarily capture all spatial correlations that are relevant for distinguishing and improving the task performance when faced with acquisition-related DS.

The FoE-CNN-PCA model

We consider additional expert functions that encode spatial correlations at the layer with the largest receptive field. To do so, we use PCA [231]↑ [232]↑. For all the training images, we extract the last layer features, $f_{cL}(z)$. Next, for each channel of $f_{cL}(z)$, we extract $r \times r$ patches with stride d. We carry out PCA of these patches and save the first G principal components. Now, for each channel c, we compute the PCA coefficients, v, for all extracted patches of all training images. The functions that output the PCA coefficients are considered the additional experts. We compute subject-wise 1D PDFs in each principal dimension, $p^s_{cg}(v)$, where $c = 1, 2, ...C_L$, $g = 1, 2, ...G$, $s = 1, 2, ...n_{tr}$.

PCA of active patches

For the task of image segmentation, we noticed that the marginal distributions of the features f_{cL} have two distinct modes - one corresponding to the regions of interest, and one to "background" regions, which are not relevant for the task at hand. In several segmentation applications, the background consists of many more pixels than the foreground classes combined. In such cases, PCA may be unable to find directions of variance within the foreground regions, matching marginal distributions of which may be more useful for TTA. To tackle this problem, we consider only active patches while doing PCA. Active patches are defined as those whose central pixel's predicted foreground segmentation probability is greater than a threshold τ.

6.3.6 How to Drive TTA using the FoE-CNN-PCA Model?

The principal components computed on the training images, as well as the expert PDFs of the principal coefficients are transferred to the test site. When a test image t arrives, patches of its features, $f_{cL}(z)$, are extracted, active patches are retained and the saved principal components are used to compute the corresponding expert PDFs, $p^t_{cg}(v)$. The matching of the additional PCA coefficient PDFs is included in the TTA optimization. That is, we minimize $L_{FoE-CNN-PCA}$ with respect to ϕ, where

$$\begin{aligned} &L_{FoE-CNN-PCA} = \\ &E_{p(s)}\Big[\tfrac{1}{L}\textstyle\sum_{l=1}^{L}\tfrac{1}{C_l}\sum_{c=1}^{C_l} D_{KL}(p^s_{cl}(u), p^t_{cl}(u)) \; + \; \lambda\,\tfrac{1}{C_L}\sum_{c=1}^{C_L}\tfrac{1}{G}\sum_{g=1}^{G} D_{KL}(p^s_{cg}(v), p^t_{cg}(v))\Big] \end{aligned} \tag{6.15}$$

A hyperparameter, λ, is used to weigh the contribution of the PCA experts with respect to the CNN ones.

We validated the proposed method for tackling the DS problem on two medical image analysis tasks - segmentation (Sec. 6.4) and atlas registration (Sec. 6.5).

6.4 Image Segmentation Experiments and Results

6.4.1 Datasets

We considered MRI segmentation for 5 anatomies (names of the segmented foreground classes are shown brackets) - (i) T2w prostate (whole organ), (ii) Cine cardiac (myocardium, left and right ventricles), (iii) T1w spine (spinal cord grey matter), (iv) healthy T1w brain (cerebellum gray matter, cerebellum white matter, cerebral gray matter, cerebral white matter, thalamus, hippocampus, amygdala, ventricles, caudate, putamen, pallidum, ventral DC, CSF and brain stem) and (v) diseased FLAIR brain (cerebral white matter hyperintensities). In total, we used data from 17 centers. Table 6.1 summarizes the details of all datasets. Please refer to Sec. 3.1 for a detailed description.

For datasets where the total number of images was very small, splits were created as indicated in Table 6.1, and average test scores are reported. The dataset splits were designed in such a way that we had 10 test volumes from each test distribution (except for the spine images, where the number of test volumes was 9).

6.4.2 Pre-Processing

The same pre-processing pipeline was used, as described for TTA-DAE in Sec. 5.4.2. That is, we (a) corrected bias fields from the images, (b) linearly normalized their intensities to 0-1 range, (c) removed the skulls from brain images, (d) rescaled the images to match their resolutions in the in-plane direction: $0.625mm^2$, $1.33mm^2$, $0.25mm^2$, $0.7mm^2$ and $1.0mm^2$ for prostate, cardiac,

Dataset	Center	Train / Test	N_I	$N_{tr}\|N_{vl}\|N_{ts}$
Prostate				
NCI-13	RUNMC	Train	30	15|5|10
NCI-13	BMC	Test	30	15|5|10
Promise12	UCL	Test	13	(6|2|5)x2
Promise12	HK	Test	12	(5|2|5)x2
Promise12	BIDMC	Test	12	(5|2|5)x2
Private	USZ	Test	68	48|10|10
Heart				
M&Ms	CSF	Train	50	30|10|10
M&Ms	UHE	Test	25	10|5|10
M&Ms	HVDH	Test	75	55|10|10
Spinal Cord Grey Matter				
SCGM	PM	Train	10	(5|2|3)x3
SCGM	USZ	Test	10	(5|2|3)x3
SCGM	VU	Test	10	(5|2|3)x3
SCGM	UCL	Test	10	(5|2|3)x3
Brain (Healthy)				
HCP	HCP-T1	Train	35	20|5|10
ABIDE	ABIDE-AC-T1	Test	25	10|5|10
Brain (White Matter Hyperintensities)				
WMH-17	UMC	Train	20	(10|5|5)x2
WMH-17	NUHS	Test	20	(10|5|5)x2

Table 6.1: Details of segmentation datasets for 5 anatomies. N_I refers to the total number of 3D images, and the last column refers to the training, validation and test split. For some datasets, the split is followed by x2 or x3. This refers to the number of dataset splits that were done to get a reasonable number of test images in datasets with a low N_I.

spine, brain and WMH respectively, and (e) cropped / padded zeros to have the same in-plane image size: 200x200 for the spine images and 256x256 for other anatomies. The evaluation for each test image was done in its original resolution and size.

6.4.3 Common Implementation Details for all Experiments

Network Architectures: We used the same architecture for N_ϕ and S_θ as in TTA-DAE [67]↑. N_ϕ consisted of 3 convolutional layers of kernel size 3, number of output channels 16, 16 and 1, and an expressive activation function ($act(x) = \exp -(x^2/\sigma^2)$) with a learnable scale σ for each channel. S_θ followed a U-Net [205]↑like encoder-decoder structure with skip connections, and batch normalization layers following each convolutional layer. The ReLU activation function was used in S_θ.

6.4.4 List of Experiments and Specific Implementation Details

For each anatomy, we used the institution in the first row (for that anatomy) in Table 6.1 as the training distribution, and the remaining institutions as separate test distributions. In this setup, we carried out the following experiments:

(I) Baseline

We trained a CNN (N_ϕ + S_θ) using labelled images from the training distribution. The supervised training was done by minimizing the Dice loss [209]↑using an Adam optimizer with a learning rate of 0.001 and a batch size of 16. The optimization was run for 30000 iterations, and the model selection criterion was the average Dice score on the validation dataset.

(II) Strong Baseline - Data Augmentation

Several domain generalization methods have been proposed to tackle acquisition-related DS in medical image analysis. From our experiments in the previous chapter, we found that stacked data augmentations [101]↑is an effective and general DG approach. The implementation details were the same as in [67]↑: for every image in a training batch, each transformation (translation, rotation, scaling, elastic deformations, gamma contrast modification, additive brightness and additive Gaussian noise) was applied with probability 0.25. This functioned as a strong baseline, the performance of which we sought to improve with the proposed TTA approach.

(III) Benchmark

The best performance on images from a test distribution can be achieved by training a new model in a supervised manner, using a separate set of labelled images from the test distribution. As some of the datasets contained only a small number of images to start with, we instead used a transfer learning benchmark - that is, the model trained on the training distribution (with data augmentation) was fine-tuned using labelled images from the test distribution. The fine-tuning was done with the Adam optimizer for 5000 iterations, with a learning rate of 0.0001 and batch size of 16. This model served as the benchmark.

(IV) Test-Time Adaptation Methods

We compared the proposed approach (TTA-FoE-CNN-PCA) with three existing TTA works: TTA-Entropy-Min [123]↑, TTA-DAE [67]↑and TTA-AE [68]↑.

Common details for all TTA methods: Using the 'strong baseline' model as the starting point, TTA was run for N_{tta} epochs for each test subject. In each epoch, averaged gradients over batches of size b_{tta} were used to update the network parameters with a learning rate of lr_{tta}. N_{tta} was set to 200 for the healthy brain dataset (due to its high through-plane size) and to 1000 for all other datasets. b_{tta} was set to 8 for all datasets except SCGM, where it was set to 2 as some images had less than 8 slices.

(IV.A) TTA-DAE [67]↑

A 3D denoising autoencoder was trained in the space of segmentation labels, using the same corruption distribution as proposed in the original paper. Similar to the original implementation, healthy brain segmentations were downsampled in the through-plane direction by a factor of 4, to overcome memory issues. lr_{tta} was set to 0.001.

(IV.B) TTA-EM [123]↑

The normalization module, N_ϕ, was adapted for each test subject, with lr_{tta} as 0.0001, to minimize the average pixel-wise entropy computed over all prediction classes.

(IV.C) TTA-AE [68]↑

Instead of adapting N_ϕ, adaptor modules A^x, A^1, A^2 and A^3 were introduced and adapted for each test subject as was done in the original article. We experimented with different settings of [68] so as to get the best results for the datasets used in our experiments (Appendix 6.7.3). The architectures of the adaptors were kept the same as proposed in [68]↑, with one change: the instance normalization layers in A_X were discarded as they lead to instability during TTA. Two other changes were done to further improve the performance and stability: (a) average gradients over all batches in a single TTA epoch were used for the TTA updates (as described in Sec 3.5 in [67]↑) and (b) the lr_{tta} was set to 0.00001. Five 2D autoencoders (AEs) (with the same architectures as in [68]↑) were trained and the weight of the orthogonality loss, λ_{orth}, was set to 1.0, as done in [68]↑. We observed that driving the TTA using losses from two AEs (at the input and output layers) provided better performance than using all 5 AEs. With these modifications, TTA-AE worked in a stable manner, without resorting to early stopping as done in [68].

(IV.D) TTA-FoE-CNN-PCA

At the end of the 'strong baseline' training, the FoE-CNN model was constructed by computing 1D PDFs for all channels of all layers of S_θ, for each training subject. For the chosen architecture of S_θ, this amounted to 704 channels. Any zero-padding done to the images in the pre-processing step was ignored while computing the expert PDFs. As the PDFs were approximated as Gaussians, two parameters were stored per PDF.

For computing the additional expert PDFs of the FoE-CNN-PCA model, the following steps were followed: (a) For all training images, features from the last layer of S_θ were extracted (from here, a 1x1 convolutional layer provided the segmentation logits). In the chosen architecture, these features were of the same spatial dimensions as the images and had $C_L = 16$ channels. (b) For each channel in these features, patches of size $r \times r = 16 \times 16$ were extracted with stride $d = 8$. (c) From these, only active patches (that is, patches whose central pixel's predicted foreground probability was greater than $\tau = 0.8$) were retained. As CNNs typically make high confidence predictions, this step is likely to be insensitive to the exact value of τ. To obtain a comparable number of active patches to other anatomies, the stride d was set to 2 for the WMH images, where the foreground size was particularly small. (d) PCA was done using the active patches of all training images, and the first $G = 10$ principal directions were identified. (e) Finally, 1D PCA expert PDFs were computed similar to the 1D CNN expert PDFs: for all channels of the last layer of S_θ, for all principal directions, for each training subject. In total, we had $C_L \times G = 160$ PCA expert PDFs for each training subject. The hyperparameter, λ, was empirically set to 0.1 (Sec. 6.4.5), and lr_{tta} to 0.0001.

(V) Analysis Experiments

(V.A) Approximating Expert Distributions with KDEs rather than as Gaussians

In the proposed method, we approximate the individual expert distributions of the FoE model (Eqn. 6.9) as Gaussian distributions. As the expert distributions are in 1D, we also considered non-parametric estimation methods, such as kernel density estimation (KDE) [233]↑ [234]↑ [235]↑. In general, KDEs have the two important downsides. Firstly, the number of data points required to get a reliable density estimate grows exponentially with dimensionality. This is not a concern in low dimensions. Secondly, KDEs require access to the training samples to evaluate the PDF at a given test sample. Again, in low dimensions (e.g 1D), it may be feasible to evaluate and save the KDE over the entire domain of interest when one has access to the training samples. Thus, the training samples are no longer required at test time. Accordingly, we compute

$$p_{cl}(u) = \tfrac{1}{N_z} \textstyle\sum_z \tfrac{1}{N_{xl} * N_{yl}} \textstyle\sum_i \tfrac{1}{\sqrt{2\pi}} \exp\left(-\alpha \, ||u - u_i||_2^2\right)) \tag{6.16}$$

Being more expressive than Gaussians, KDEs can potentially capture higher-order moments of the expert distributions - thus leading to more accurate distribution matching and better TTA performance.

Implementation-wise, when the 1D PDFs were estimated as Gaussians, the KL-divergence could be computed in closed form. When KDEs are used, we numerically compute the integral in the KL-divergences using Riemann sums.

(V.B) Effect of the weighting between the CNN and the PCA experts

The effect of the weighting parameter, λ, in Eqn. 6.15, was empirically analyzed for the 5 test distributions of the prostate segmentation experiment.

6.4.5 Results

The following points can be inferred from the quantitative results of our segmentation experiments (Table 6.2).

(I) Baseline

The baseline demonstrates the DS problem. The difference between the Dice scores on the training and test distributions is sometimes as high as 60 Dice points; a model that provides almost perfect segmentations on the training distribution can potentially provide completely un-usable segmentations on images from a test distribution that corresponds to a different hospital.

(II) Strong Baseline

The strong baseline (data augmentation [101]↑) helps vastly. It is much more robust to DS than the baseline - in some cases, the performance jump is as high as 50 Dice points. These results corroborate numerous similar findings in the current literature. Given the generality and effectiveness of the approach, we believe it is imperative that works studying DS robustness in CNN-based medical image segmentation should include stacked data augmentation during training.

(III) Benchmark

A gap to the benchmark still remains - in most cases, heuristic data augmentation falls short of rivalling the performance of supervised fine-tuning.

(IV.A) TTA-DAE

Among TTA methods, [67]↑provides the best performance for the most number of cases. However, it also leads to a drop of 5 and 8 Dice points from the two spine datasets, and fails to improve performance for the WMH dataset. For the latter case, we speculate that this reflects the inability of the DAE to learn a reliable shape prior.

(IV.B) TTA-EM

Test Method	UCL	HK	BIDMC	BMC	USZ	UHE	HVHD	USZ	VU	UCL	AC	NUHS
	Prostate					Cardiac		Spine			Brain	WMH
	Supervised Learning on Training Distribution											
Baseline	0.50	0.68	0.29	0.28	0.67	0.86	0.38	0.61	0.82	0.79	0.69	0.00
	Domain Generalization											
Strong baseline [101]↑	0.77	0.82	0.62	0.77	0.76	0.85	0.80	0.67	0.84	0.88	0.76	0.37
	Test Time Adaptation											
Entropy Min. [123]↑	0.77	0.81	$0.68^{\triangle}$	0.77	$0.80^{\blacktriangle}$	0.85	$0.80^{\triangle}$	0.67	0.84	0.88	$0.81^{\blacktriangle}$	$0.36^{\blacktriangledown}$
DAE [67]↑	$\mathbf{0.84}^{\blacktriangle}$	$\mathbf{0.84}^{\triangle}$	$\mathbf{0.75}^{\blacktriangle}$	$\mathbf{0.81}^{\triangle}$	$\mathbf{0.82}^{\blacktriangle}$	$\mathbf{0.87}^{\blacktriangle}$	0.81	**0.69**	$0.80^{\triangledown}$	0.80	$\mathbf{0.82}^{\blacktriangle}$	0.37
AE [68]↑	0.78	0.83	$0.51^{\triangledown}$	**0.79**	0.79	$0.86^{\blacktriangle}$	0.80	$\mathbf{0.69}^{\triangle}$	$0.84^{\triangle}$	$0.88^{\triangle}$	$0.78^{\blacktriangle}$	$0.24^{\blacktriangledown}$
FoE-CNN [236]↑	0.78	$0.77^{\triangledown}$	0.64	0.76	0.76	0.86	$\mathbf{0.82}^{\blacktriangle}$	0.68	$\mathbf{0.85}^{\triangle}$	$\mathbf{0.89}^{\triangle}$	$0.79^{\blacktriangle}$	$0.24^{\blacktriangledown}$
FoE-CNN-PCA (Ours)↑	0.79	0.81	$0.73^{\triangle}$	0.75	0.78	0.85	$\mathbf{0.82}^{\blacktriangle}$	0.68	$0.83^{\triangledown}$	0.88	$0.79^{\blacktriangle}$	$\mathbf{0.42}^{\blacktriangle}$
	Transfer Learning											
Benchmark	0.80	0.85	0.82	0.83	0.84	0.88	0.83	0.78	0.85	0.90	0.88	0.77

Table 6.2: Dice scores (averaged over all foreground labels and all test subjects) for the segmentation test-distribution datasets. In each column, the highest Dice score among the TTA methods has been highlighted. The Dice scores for test images from the training distribution are: (a) for the baseline: RUNMC 0.86, CSF: 0.82, PM: 0.88, HCP: 0.87, UMC: 0.71, (b) for the strong baseline: RUNMC 0.91, CSF: 0.83, PM: 0.89, HCP: 0.87, UMC: 0.72. Results for the NUHS dataset are mean values over 4 runs. Paired permutation tests were done to measure the statistical significance of the improvement or degradation caused by each TTA method over the strong baseline. $^{\triangle}$ ($^{\triangledown}$) and $^{\blacktriangle}$ ($^{\blacktriangledown}$) indicate improvement (degradation) with p-value less than 0.05 and 0.01, respectively. The stricter significance test (p-value 0.01) was done to counter the multiple comparison problem [237]↑.

Entropy minimization-based TTA [123]↑requires construction of no additional models to capture the training distribution traits; yet, it provides performance improvement in several cases. Also, unlike other works [140]↑, we do not observe the problem that the entropy minimization leads to all pixels being predicted as the same class. This might have been due to the limited adaptation ability provided by N_ϕ.

(IV.C) TTA-AE

Autoencoder-based TTA [68]↑provides performance improvement in several cases. However, it also leads to a drop of 12 and 26 Dice points for the prostate BIDMC and the WMH dataset, respectively.

(IV.D) TTA-FoE-CNN-PCA

The last two rows of Table 6.2 show the added benefit of the PCA experts to the FoE model.

As compared to the strong baseline, the proposed FoE-CNN-PCA based TTA improves performance for 7 and retains performance for 2 out of the 12 test distributions. In particular, the proposed method shows promising performance gains in cases where the competing methods falter substantially (e.g.

prostate BIDMC and WMH). Further, for the 3 test distributions where the method leads to a performance drop, the drop is relatively small: 3, 1 and 1 Dice points. We claim that this illustrates the **stability** of the proposed TTA method and validates our initial hypothesis - FoE-based TTA improves performance in the face of acquisition-related DS in medical imaging, while itself being substantially more robust to the DS shift problem that other priors such as the DAE [67]↑or the AE [68]↑may be vulnerable to.

Additionally, the proposed method provides the best performance for the task

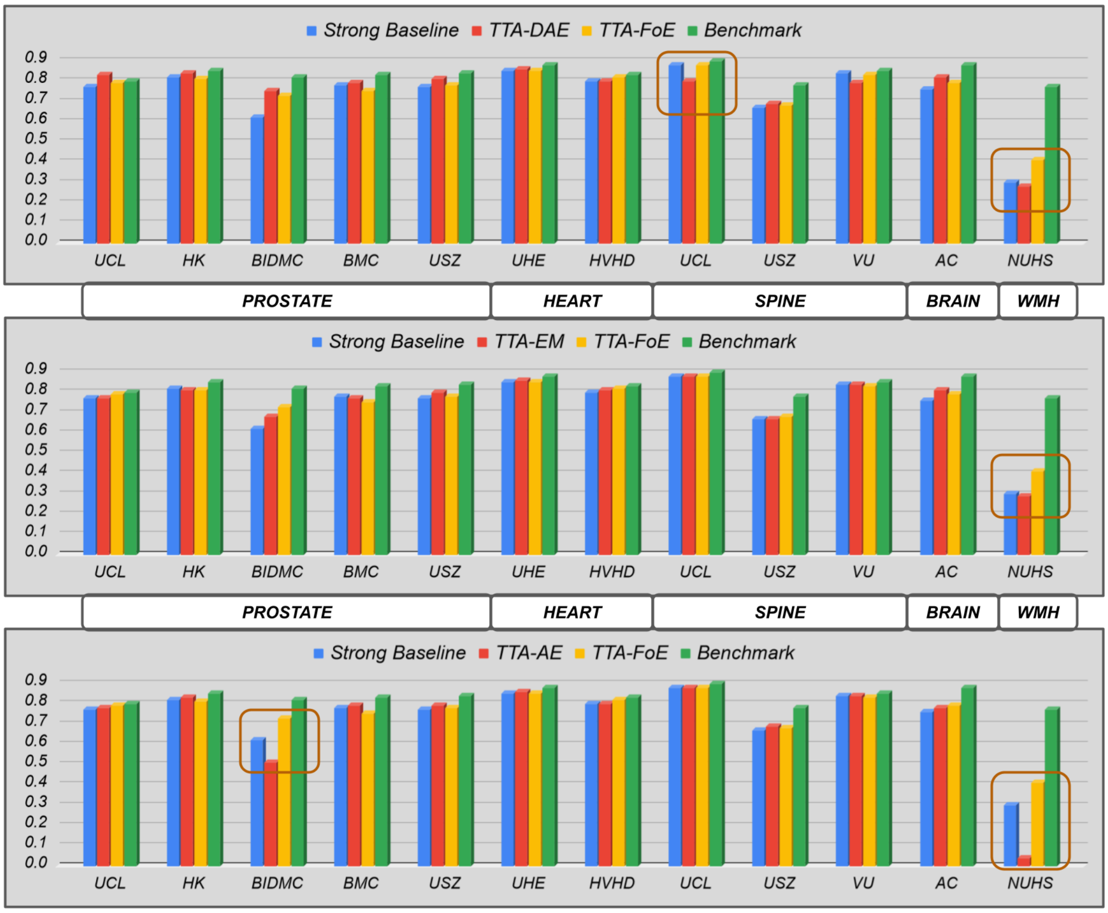


Figure 6.3: Graphical comparison of TTA methods. Each of the three subplots compares the performance of TTA-FoE-CNN-PCA with one competing TTA method from the literature: TTA-DAE (top), TTA-EM (middle) and TTA-AE (bottom). Additionally, the strong baseline and benchmark results are shown in each subplot. All TTA methods perform similarly for most datasets - after carefully tuning hyper-parameters for each method. Yet, the proposed TTA method is more stable (does not lead to drastic performance drop in any dataset) and more general (provides performance improvement for the lesion segmentation problem, while none of the existing methods manage to do so). Brown boxes highlight such scenarios.

of WMH segmentation - indicating its superiority in cases where CNN-based helper modules such as DAEs [67]↑may be unable to learn appropriate shape priors. Notably, all competing methods from the literature fail to improve DS robustness for the lesion segmentation experiment; the proposed TTA-FoE-CNN-PCA is the only approach that shows promising results in this challenging scenario. We claim that this demonstrates **generality** of the proposed approach over previously existing methods. Fig. 6.4 shows the evolution of the predicted segmentation for the WMH test subject, over the course of TTA-FoE-CNN-PCA iterations. The prediction becomes better as the expert distributions of the test image better overlap with the corresponding expert distributions for training subjects.

A graphical comparison of the different TTA methods is shown in Fig. 6.3, which visually corroborates the discussion above.

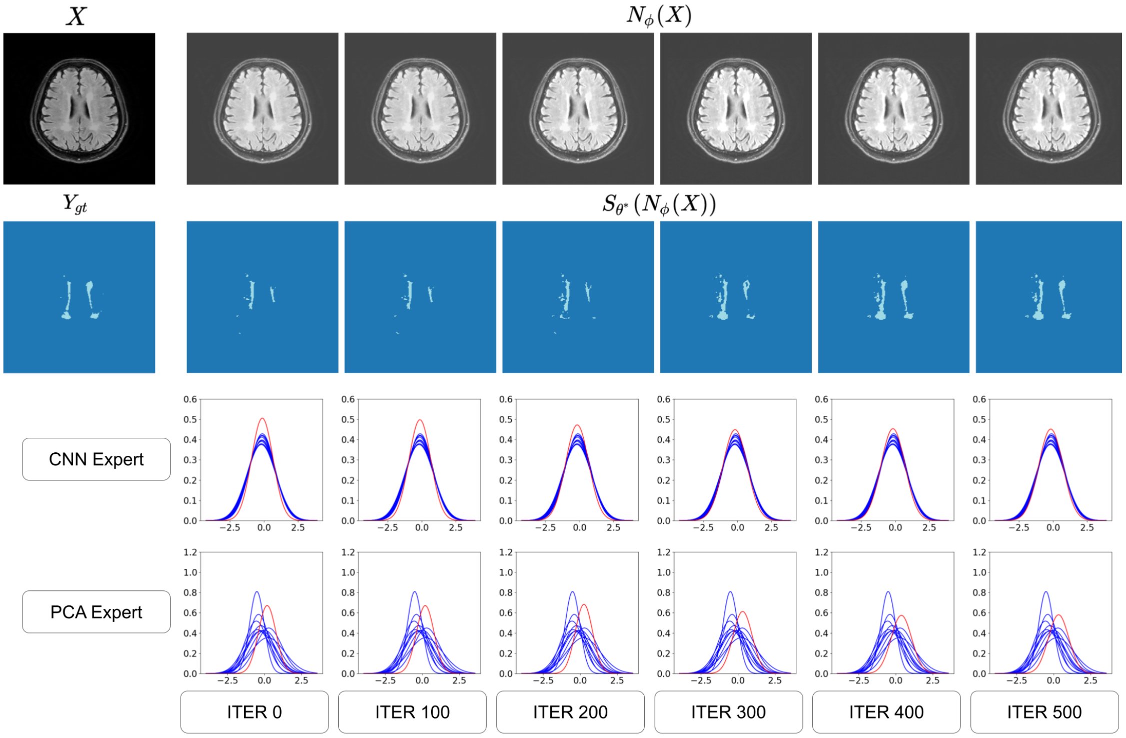
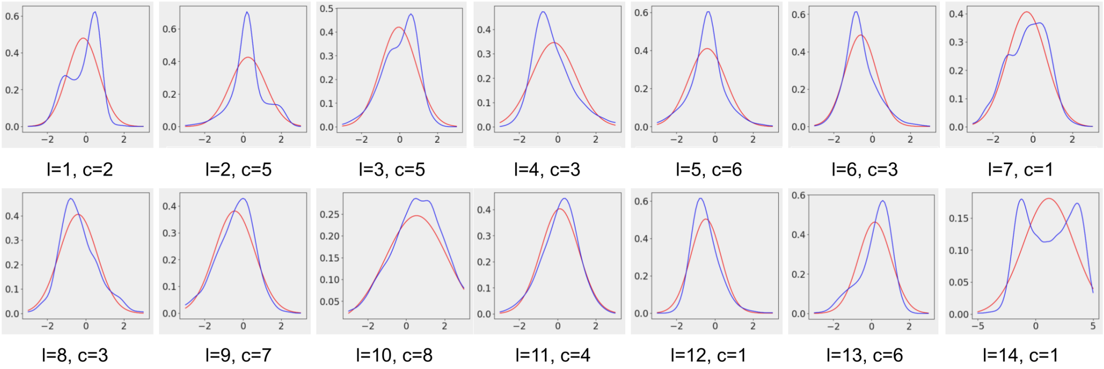


Figure 6.4: Evolution during TTA iterations (for WMH segmentation in a NUHS test distribution subject) of (from top to bottom): (i) normalized images, (ii) predicted segmentations, overlap between training and test distributions of (iii) a CNN expert and (iv) a PCA expert. For the last two rows, the distributions corresponding to different training subjects are shown in blue, and that corresponding to the test image is shown in red.

(V) Analysis Experiments

(V.A) Approximating Expert Distributions with KDEs rather than as Gaussians

Comparing the KDEs v/s Gaussian approximations (Fig. 6.5), we observed that the actual distributions do not differ substantially from their Gaussian approximations. This is also reflected in the TTA results in Table 6.3 - performance of the proposed method is very similar for both estimates of expert distributions.


l=1, c=2 l=2, c=5 l=3, c=5 l=4, c=3 l=5, c=6 l=6, c=3 l=7, c=1

l=8, c=3 l=9, c=7 l=10, c=8 l=11, c=4 l=12, c=1 l=13, c=6 l=14, c=1

Figure 6.5: Comparison of KDEs v/s Gaussian approximations (corresponding to a single prostate RUNMC training subject) for modeling the channel PDFs of different layers of the trained segmentation network. $l = 14$ is the last-but-one layer of the network. From here, a 1x1 convolution gives the segmentation logits. In each layer (l), the channel (c) with the visually most-non-gaussian KDE is chosen for visualization. With this choice, some non-Gaussianity is observed in the initial and final layers, while the layers in the middle of the segmentation CNN has highly Gaussian marginal distributions.

Test Method	UCL	HK	BIDMC	BMC	USZ
			TTA-FoE-CNN-PCA		
Gaussian	0.79	0.81	0.75	0.75	0.78
KDE	0.79	0.81	0.74	0.76	0.78

Table 6.3: Effect of approximating 1D distributions of the FoE model with Gaussians v/s kernel density estimation (KDE). Both approximations lead to very similar TTA performance. Fig. 6.5 provides visual justification of this observation - the 1D distributions of CNN as well as the PCA experts are sufficiently well approximated with Gaussians.

(V.B) Effect of the weighting between the CNN and the PCA experts

Results of this hyper-parameter tuning are shown in Table 6.4. The introduction of PCA experts with $\lambda = 0.1$ improves TTA performance for 4 of the 5 prostate datasets. However, increasing λ to 1.0 leads to performance decrease in 4 of the 5 datasets. Based on these results, we choose $\lambda = 0.1$ for all datasets of all anatomies.

Test Method	UCL	HK	BIDMC	BMC	USZ
			TTA-FoE-CNN		
$\lambda = 0.0$	0.78	0.77	0.64	0.76	0.76
			TTA-FoE-CNN-PCA		
$\lambda = 0.1$	**0.79**	0.81	**0.75**	0.75	**0.78**
$\lambda = 1.0$	0.77	**0.82**	0.74	0.74	0.77

Table 6.4: Effect of the weighting parameter between the CNN and PCA experts in TTA-FoE-CNN-PCA. Based on these results, we choose $\lambda = 0.1$ for all datasets of all anatomies.

6.5 Image Registration Experiments and Results

Next, we checked if the proposed method can tackle acquisition-related DS in another task of high practical importance - registration of brain scans with an atlas.

Registration CNN Setup: The registration CNN is set up as follows. (Ideally, such registration would be done in 3D. However, to avoid memory issues in 3D CNNs, we conduct experiments in a 2D setup. We believe that this still serves as credible evidence of the method's applicability in this task.) Let A be an atlas and X be the image. Let A_s and X_s be the corresponding segmentation labels. We treat A as the moving image and register it to x, the fixed image. X is first passed through the normalization module, N_ϕ, to obtain a normalized image, Z. Z and A are concatenated and passed through a deep CNN, S_θ, which outputs a velocity field V_0. V_0 is exponentiated via a squaring-and-scaling layer [238]↑to obtain a diffeomorphic deformation field, Φ.

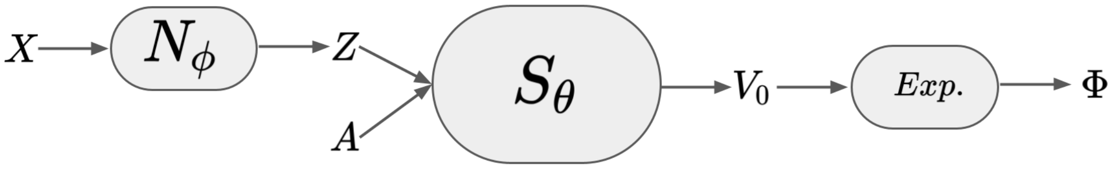


Figure 6.6: Setup of our registration experiments. $Exp.$ denotes an squaring and scaling exponentiation layer.

The Dice loss between the warped moving segmentation, $A_s \odot \Phi$, and X_s is used for training N_ϕ and S_θ. The image-specific normalization module, N_ϕ is adapted, for each test image.

6.5.1 Datasets

We used HCP [186]↑T1w images as those from the training distribution and ABIDE-STANFORD (AS) [187]↑↑ and OASIS [196]↑as two test distributions. We used the atlas provided by [197]↑↑. Example images are shown in Fig. 6.7.

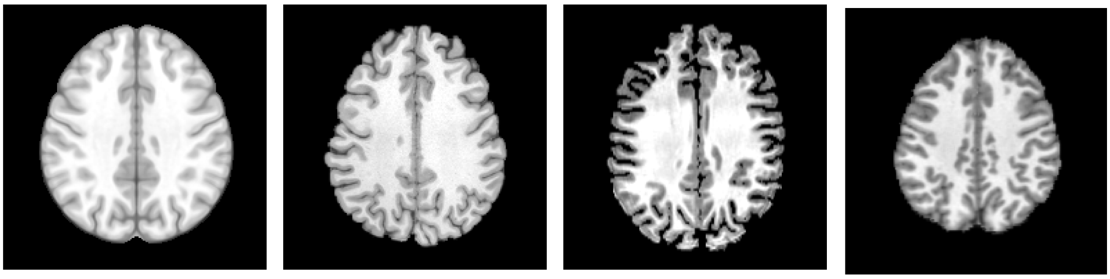


Figure 6.7: From left to right: a 2D slice from the atlas and example slices from three datasets: HCP, ABIDE-STANFORD (AS) and OASIS.

6.5.2 Common Implementation Details for all Experiments

All images were re-sampled to an isotropic $1\ mm^3$ resolution. Upon visual inspection, the axial slices of the atlas, the HCP and OASIS datasets were roughly aligned in the through plane direction, while the AS volumes were shifted by 10 slices. After accounting for this, we extracted the central 40 axial slices from all volumes. We used 3-label (background, white matter, grey matter) Freesurfer [194] segmentations for HCP, AS and expert segmentations for the atlas and OASIS.

6.5.3 List of Experiments and Specific Implementation Details

(I) Baseline, (II) Strong baseline and (III) Benchmark

Similar to Sec. 6.4.3, the baseline is supervised learning on the training distribution, strong baseline is supervised learning with extensive data augmentation [101]↑and benchmark is transfer learning (fine-tuning using on the test distribution).

(IV) Test-Time Adaptation Methods

Among the TTA methods, we note that TTA-EM [123]↑and TTA-DAE [67]↑are not applicable for the image registration experiments.

TTA-EM [123]↑can only be applied in cases where S_θ outputs a probability distribution over a fixed number of classes; it is unclear how to extend this for regression problems.

TTA-DAE [67]↑requires a denoising autoencoder to be trained with corruption patterns that are expected at test time. Designing such corruptions for the registration task is non-trivial.

Thus, we compare the proposed method **TTA-FoE-CNN-PCA** with **TTA-AE** [68]↑. As with segmentation, we carry out TTA for the base network trained with data augmentation (strong baseline).

6.5.4 Results

Quantitative results of our registration experiments are shown in Table 6.5.

(I) Baseline, (II) Strong baseline and (III) Benchmark

The baseline shows that the DS problem exists for the registration problem as well. Further, data augmentation [101] provides substantial gains for registration - demonstrating the wide generality of this approach for tackling DS. The strong baseline almost matches the benchmark performance for the OASIS dataset, while a gap in performance exists for the ABIDE-STANFORD (AS) dataset.

(IV) Test-Time Adaptation Methods

Both TTA-FoE-CNN-PCA and TTA-AE [68]↑improve the performance for the AS dataset and retain it for the OASIS dataset. This shows the applicability of both approaches to multiple tasks. We argue that such across-task applicability is an important strength of the proposed method.

Fig. 6.8 shows subject-wise results for the AS dataset. Like for the segmentation experiments, it can be seen that both TTA methods perform similarly for most subjects. For one subject, however, TTA-AE leads to performance degradation as compared to even the baseline. Notably, such degradation does not happen for any subject with the proposed method. This may be seen as further evidence of the proposed method's stability. We note, however, more experimentation may be required to validate this claim for the image registration task, and for other tasks.

6.6 Discussion

In this chapter, we proposed a task-agnostic TTA method, TTA-FoE, for improving robustness to acquisition-related DS in medical image analysis. TTA-FoE is motivated by the notion that per-image adaptability is crucial for developing robust medical image analysis tools, as introduced in the previous

Test Method	HCP	AS	OASIS
Baseline	0.847	0.751	0.864
Strong baseline [101]↑	0.843	0.786	0.873
Benchmark	-	0.821	0.883
TTA-DAE [67]↑	-	N/A	N/A
TTA-EM [123]↑	-	N/A	N/A
TTA-AE [68]↑	-	0.795	0.868
TTA-FoE-CNN-PCA	-	0.795	0.870

Table 6.5: Dice scores (averaged over all foreground labels and all test subjects) for the registration experiments. The two TTA methods that can be applied to this task perform similarly well to one another, while the other two TTA methods cannot be applied to the image registration task.

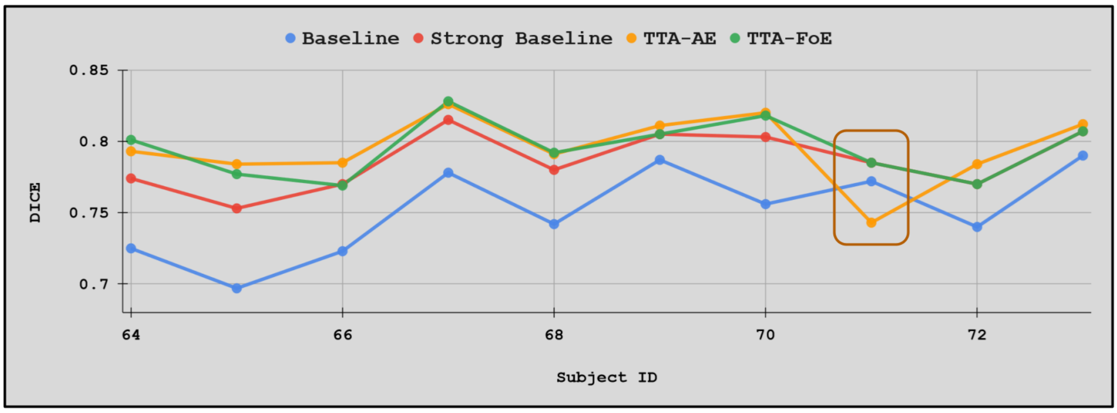


Figure 6.8: Dice scores for individual subjects of the AS dataset. Overall, both TTA methods perform similarly well for most subjects. However, the brown box highlights one subject, where TTA-AE leads to performance degradation as compared to both the baseline as well as the strong baseline. Such degradation does not occur with the proposed method.

chapter. It also follows the method of the previous chapter (TTA-DAE) in that it adapts a shallow normalization sub-network for each test image.

The procedure to achieve the TTA in TTA-FoE improves upon that of TTA-DAE in two respects. (1) TTA-FoE is carried out by matching distributions of the test and training images in the output space of normalization sub-network. The distributions in this space are modelled using a Field-of-Experts (FoE) formulation. By relying on simpler distributions of a large number of task-specific 1D projections of normalized images, the FoE model is less vulnerable the DS robustness problem than CNN-based helper models that are used to drive TTA in TTA-DAE [67]↑and TTA-AE [68]↑. The improved robustness of the TTA helper model manifests as improved TTA stability - the proposed method provides performance gains in a stable manner, while existing methods in the literature lead to performance degradation over the baseline in several instances. (2) TTA-DAE relies on a denoising autoencoder in the output space, training which might be non-trivial for dense prediction tasks other than segmentation and infeasible for sparse predictions tasks such as classification. The distribution matching procedure in TTA-FoE, on the other hand, is task-agnostic. Thanks to the increased generality, the method could be used to improve DS robustness for the tasks of image registration as well as image segmentation.

In the following paragraphs, we discuss the strengths and limitations of TTA-FoE as well as some of design choices, the reasoning for these choices and avenues that could potentially further improve TTA-FoE performance.

6.6.1 Strengths of TTA-FoE

1. **Lesion segmentation performance**: All existing TTA methods failed to tackle the DS robustness problem for lesion datasets. Furthermore, 3 out of the 4 existing methods lead to statistically significant performance degradation over the strong baseline. In particular, TTA-DAE, which shows strong performance for healthy tissue segmentation, fails to improve performance for lesions due to the difficulty in learning appropriate shape priors. The proposed method provided substantial as well as statistical significant performance improvement in this challenging scenario.

2. **Applicability to multiple tasks**: Our experiments indicate that the proposed method can, in principle, be applied to multiple tasks. Such generality is an important asset; the DS problem is likely to occur in all medical image analysis tasks.

3. **Generalization of previous works**: This work makes the novel contribution of casting the marginal distribution matching idea in a Field-of-Experts formulation. This observation allows us view several recent works [131]↑, [132]↑, [239]↑, [236]↑as instances of our general framework, and enables us to build on these works by introducing additional expert functions in the form of principle loadings of feature patches.

6.6.2 Limitations of TTA-FoE

1. **Performance on healthy tissue segmentation is not as good as TTA-DAE**: Although the proposed method improves performance of the strong baseline

in a large number of the test datasets, methods specifically designed for image segmentation often outperform the more general method developed in this work.

2. **Matching the distribution of individual experts rather than the full FoE distribution**: An important relaxation in TTA-FoE is between Eqn. 6.12 and Eqn. 6.13. Eqn. 6.12 seeks to match the full FoE distribution between test and training images. However, this is not possible as the computation of the normalization constant $\mathcal{C}$ is intractable. Thus, we instead carry out the relaxed optimization, as shown in Eqn. 6.13 - minimizing divergence between the distributions of individual experts. It is unclear if the relaxed optimization is theoretically guaranteed to converge, or if the alignment of individual experts may compete with one another. In practice, we observe the optimization to converge for all the test images, across all test distributions and anatomical regions. We believe that this behaviour could have been aided by the initial closeness of the individual expert distributions. Thus, the proposed TTA method works well for small DS (due to changing scanners or acquisition protocol parameters within the same imaging modality), but may not be suitable for large DS (for instance, across imaging modalities).

6.6.3 Avenues for further exploration

1. **Choice of expert functions of the FoE Model**: In initial product-of-experts [228]↑, [229]↑and field-of-experts [226]↑works, the experts are parameterized and learned from data, such that the probability model assigns high likelihood values to the true data - for example, using algorithms such as contrastive divergence. Further, parameters of the expert PDFs are also learned from data. In contrast, in this work, we used two types of experts - (1) the task-specific convolutional filters learned in the segmentation or registration CNN and (2) projections onto principal components of patches in the last layer of the segmentation or registration CNN. Thus, we used task-specific experts, and only learned the parameters of the expert PDFs from data. In other words, we aligned the test and training normalized images, in terms of their projections that are the most relevant for the task CNN to perform the task at hand. Such a task-specific probability model could be augmented with learned experts, as proposed in earlier works [228]↑, [229]↑, [226]↑. The extended model would potentially capture further projections of the normalized images, apart from the task-specific projections considered in this work. It is unclear if alignment along such directions between test and training images would further improve TTA performance; we defer this analysis to future work.

2. **Choice of the divergence measure to be minimized for TTA**: We minimize the KL-divergence between expert distributions. Other divergence measures may also be considered. For instance, in concurrent work, [236]↑minimize a symmetric version of the KL divergence. Leveraging the low dimensionality of the expert outputs, even divergence measures that cannot be computed in closed form, may be easy to compute numerically.

6.7 Appendix

6.7.1 Approximating KL-divergence minimization of the full FoE model with KL-divergence minimization of individual expert distributions

We show this analysis for Product of Experts (PoEs). It also holds for FoEs, which are a specific instance of the PoEs formulation. Consider PoE models for the source and target domain normalized images.

$$p^s(z) = \frac{\hat{p}^s(z)}{\mathcal{C}_s}, \quad \mathcal{C}_s = \int_z \hat{p}^s(z) dz, \quad \hat{p}^s(z) = \prod_{j=1}^{J} p_j^s(u_j), \quad u_j = f_j(z)$$

$$p^t(z) = \frac{\hat{p}^t(z)}{\mathcal{C}_t}, \quad \mathcal{C}_t = \int_z \hat{p}^t(z) dz, \quad \hat{p}^t(z) = \prod_{j=1}^{J} p_j^t(u_j), \quad u_j = f_j(z)$$

Here, we explicitly show the subscript j in variables u to indicate that different experts have different 1D co-domains. Now, consider KL-divergence minimization between these distributions:

$$min_\phi \; D_{KL}(p^s(z), p^t(z)) \rightarrow min_\phi \int_z p^s(z) \; \log \; \frac{p^s(z)}{p^t(z)} \, dz$$

$$\rightarrow min_\phi \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \; \log \; \frac{\mathcal{C}_t}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \; \log \; \frac{\mathcal{C}_t}{\mathcal{C}_s} \, dz + \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \; \log \; \frac{\mathcal{C}_t}{\mathcal{C}_s} + \int_z \frac{\hat{p}^s(z)}{\mathcal{C}_s} \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

Note that during TTA, ϕ is fixed for computing the source-domain distribution, while is variable for computing the target-domain distribution. Thus, ignoring the 'source-domain-only' terms, the minimization can be stated as follows:

$$\rightarrow min_\phi \; \log \; \mathcal{C}_t + \int_z \hat{p}^s(z) \; \log \; \frac{\hat{p}^s(z)}{\hat{p}^t(z)} \, dz$$

$$\rightarrow min_\phi \; \log \; \mathcal{C}_t + \\ \textstyle\int_{u_1, u_2, ..u_J} \; \prod_{j=1}^{J} p_j^s(u_j) \; \log \; \frac{\prod_{j=1}^{J} p_j^s(u_j)}{\prod_{j=1}^{J} p_j^t(u_j)} \, du_1 du_2 ... du_J$$

$$\rightarrow min_\phi \; \log \; \mathcal{C}_t + \\ \textstyle\sum_{j=1}^{J} \; \int_{u_1, u_2, ..u_J} \; \prod_{j=1}^{J} p_j^s(u_j) \; \log \; \frac{p_j^s(u_j)}{p_j^t(u_j)} \, du_1 du_2 ... du_J$$

$$\rightarrow min_\phi \; \log \; \mathcal{C}_t + \sum_{j=1}^{J} \int_{u_j} p_j^s(u_j) \; \log \; \frac{p_j^s(u_j)}{p_j^t(u_j)} \, du_j$$

As the normalization constant $\mathcal{C}_t$ is intractable, we ignore it in our optimization:

$$\approx min_\phi \sum_{j=1}^{J} \int_{u_j} p_j^s(u_j) \log \frac{p_j^s(u_j)}{p_j^t(u_j)} du_j$$

$$\rightarrow min_\phi \sum_{j=1}^{J} D_{KL}(p_j^s(u_j), p_j^t(u_j))$$

6.7.2 How to incorporate information from multiple training subjects?

Consider the KL-divergence between the expected distribution over all training subjects and the distribution of the test subject. For simplicity of notation, let us consider only one 1D expert's distribution.

$$\begin{aligned}
&D_{KL}\big(E_{p(s)}[p^s(u)], p^t(u)\big)\\
&= \int_u \Big(\int_s p(s)p^s(u)ds\Big)\, log\, \frac{\int_s p(s)p^s(u)ds}{p^t(u)}\, du\\
&= \int_s p(s)\Big(\int_u p^s(u)\, log\, \frac{\int_s p(s)p^s(u)ds}{p^t(u)}\, du\Big) ds\\
&= \int_s p(s)\Big(\int_u p^s(u)\, log\, \frac{\int_s p(s)p^s(u)ds}{p^t(u)}\, \frac{p^s(u)}{p^s(u)}\, du\Big) ds\\
&= \int_s p(s)\Big(\int_u p^s(u)\, log\, \frac{\int_s p(s)p^s(u)ds}{p^s(u)}\, du + \int_u p^s(u)\, log\, \frac{p^s(u)}{p^t(u)}\, du\Big) ds\\
&= \int_s p(s)\Big(\int_u p^s(u)\, log\, \frac{E_{p(s)}[p^s(u)]}{p^s(u)}\, du + \int_u p^s(u)\, log\, \frac{p^s(u)}{p^t(u)}\, du\Big) ds\\
&= -\ \mathbb{E}_{p(s)}[D_{KL}(p^s(u), E_{p(s)}[p^s(u)])] + \mathbb{E}_{p(s)}[D_{KL}(p^s(u), p^t(u))]\\
&\leq \mathbb{E}_{p(s)}[D_{KL}(p^s(u), p^t(u))]
\end{aligned}$$

6.7.3 TTA-AE variants

[68] propose a autoencoder-based method for TTA. We made some minor changes in their method to get optimal results on the datasets used in our experiments. We did this analysis for 5 prostate segmentation test distributions, and used the optimal settings for the other datasets.

Architecture: In the proposed method, the adaptable module, N_ϕ is trained on the training distribution and further adapted for each test image. In contrast, [68] introduce 4 adaptors, A^x, A^1, A^2, A^3, as different layers in the task CNN directly at test time. A^1, A^2, A^3 are initialized to be identity mappers, while A^x is randomly initialized. In our experiments, we found that the randomly initialized A^x (with the same architecture as in [68]) substantially altered the image intensities before any TTA iterations were done. Due to this, the Dice scores at the start of TTA iterations dropped to almost 0, and could not be recovered by the TTA. We could resolve this with the help of two changes to the architecture of A^x: (i) Instead of initializing the convolutional weights with mean 0, we initialize with mean as the inverse of number input channels and variance as proposed in [240], (ii) we removed instance normalization layers from A^x. The initial Dice scores (TTA epoch 0) were now reasonable ('Architecture' in Table 6.6), although much lower than the strong baseline. The TTA iterations improve the results, but are unable to cross the strong baseline.

Optimization: We observed that the Dice scores fluctuated heavily across the TTA iterations. After reducing the learning rate from 0.001 (used in [68]) to 0.00001 and using the gradient accumulation strategy proposed in [67], we observed improved performance ('Optimization' in Table 6.6). However, the Dice scores initially improved and then dropped after about 100 epochs, for 3 of the 5 test distributions.

Loss: Plotting the evolution of the losses of the 5 AEs: one each at the input AE^x and the output layers AE^y, and 3 at different features depths (AE^{F1}, AE^{F2}, AE^{F3}) in the task CNN, we observed that the accuracy of AE^x and AE^y correlated well with the Dice scores, while this was untrue for the feature-level AEs. Thus, we carried out TTA driven only by AE^x and AE^y. In this setting, TTA-AE provided performance improvement in a stable manner ('Loss' in Table 6.6). We used this setting for the experiments on the rest of the datasets.

Test Method	UCL	HK	BIDMC	BMC	USZ
	Domain Generalization				
Strong baseline [101]↑	0.77	0.82	0.62	0.78	0.77
	TTA-AE [68]↑Variants				
Modification in:	Details				
Architecture	Removing instance normalization in A^x				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.56	0.73	0.51	0.50	0.76
Optimization	Lower learning rate, gradient accumulation				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.78	0.74	0.50	0.71	0.65
TTA Epoch 100	0.77	0.83	0.56	0.78	0.78
TTA Epoch 1000	0.65	0.78	0.57	0.73	0.79
Loss	Using AEs only at input & output layers				
TTA Epoch 0	0.76	0.71	0.48	0.67	0.57
TTA Epoch 10	0.78	0.74	0.48	0.71	0.64
TTA Epoch 100	0.79	0.82	0.51	0.78	0.78
TTA Epoch 1000	0.78	0.83	0.50	0.79	0.79

Table 6.6: Performance of TTA-AE [68]↑variants.

Chapter 7

Discussion

In this thesis, we developed three approaches to improve robustness of deep learning methods for medical image analysis to acquisition-related distribution shifts in their inputs. Of these, one approach (chapter 4) was set in the transfer learning setting, while the other two (chapters 5, 6) were set in the test-time adaptation setting. In this chapter, we discuss the links between the proposed approaches, how they relate to concurrent developments in the literature and the potential for extending them to other types of DS than the ones considered in this thesis.

7.0.1 Which parameter subset to adapt for the test distribution?

One way of grouping the machine learning settings for tackling DS (Sec. 2.2) is as follows. The first category is of adaptation methods, consisting of Transfer learning (TL) and Test-Time Adaptation (TTA). In these methods, an initial model is trained using a labelled dataset from a training distribution(s), and subsequently adapted to suit either an entire test distribution or individual test images. The other category consists of Unsupervised Domain Adaptation (UDA) and Domain Generalization (DG). Methods in this category do not have separate training and adaptation steps. Instead, a model is trained from scratch in such a way that it is suitable for the desired test images.

Among all methods in the first category, a running theme is to adapt a subset of the model parameters for the test distribution, while sharing the remaining parameters between the training and test distributions. Following the notation used in this thesis, the former and latter sets of parameters are denoted by ϕ and θ, respectively, while Θ denotes the total set of all model parameters.

Reasons for sharing a bulk of the parameters across distributions: To the best of our knowledge, a principled approach for dividing Θ into θ and ϕ is yet to be proposed. Instead, this choice is typically dictated by a-priori knowledge of the task and DS in question. Usually, the number of parameters in the shared subset θ is much larger than those in the adapted subset ϕ, due to the following reasons. First, particularly for TL methods, the limited number of parameters in ϕ is said to prevent overfitting to the small labelled dataset in the test distribution [241]↑. Second, the large size of θ is motivated by transferring learning from the training to the test distribution, and assuming that the training and test distributions are highly related. Even from so-called large DS such as T1-weighted to T2-weighted MR images or CT to MRI, the DS can primarily be modelled as contrast transformations, which can be expressed via a relatively small number of parameters.

Parameter-splitting strategies in this thesis: All three methods developed in this thesis belong to the category of adaptation methods, according to the categorization mentioned here. For the TL method (chapter 4), we set ϕ to the batch normalization parameters throughout the CNN, while for the two TTA methods (chapters 5, 6), we set ϕ to be the parameters of a shallow normalization sub-network situation in the front of the overall task CNN. The choice in chapter 4 was motivated by previous works in the computer vision literature [200]↑, [201]↑, while the choice in the TTA chapters 5, 6 was driven by domain knowledge that acquisition-related DS manifest primarily as contrast changes. With a similar motivation, [100]↑use randomly weighted shallow CNNs to generate contrast augmentations for training robust CNNs.

Parameter splitting strategies in the literature: Other adaptation methods employ different choices for ϕ. Among TL methods, [242]↑adapt all CNN parameters for the test distribution, while [243]↑use pre-trained initial CNN layers as feature extractors, and adapt a final classification layer for the test distribution. Among TTA methods, [68]↑set ϕ to be the parameters of so-called adaptor modules, which are shallow CNNs situated at the front, as well as at multiple intermediate layers of the task CNN. [123]↑set ϕ to be the batch normalization parameters in all layers of the task CNN. We suspect that the choice of ϕ may be connected to the choice to the loss used to drive TTA, as well as the type of DS at hand. As an example, using a TTA helper model at a certain depth in a task CNN (as done in [68]↑) necessitates ϕ to contain parameters of the task CNN of preceding layers only. Thus, different choices of ϕ may be necessary for different TTA losses and for tackling different DS. An understanding of these relationships is missing from the literature, to the best of our knowledge, especially within the medical image analysis community.

Methodological approaches for selecting the parameter split: In the wider computer vision literature, [241]↑provide an empirical analysis of the transferability of different CNN layers across tasks. Recent works [244]↑, [124]↑, [76]↑offer more methodological approaches for answering this question. [124]↑propose a Bayesian framework for TTA, which allows all model parameters to be adapted for the test distribution, but vicinity to the optimal parameters for the training distribution is encouraged. [244]↑propose at a similar regularization strategy for TL. Thus, the question of parameter splitting is circumvented in these two approaches. [76]↑employ a separate routing CNN, which decides for each training image from the test distribution, which layers of the task CNN should be fine-tuned. Thus, in this transfer learning setup, the subset ϕ is different for each training image of the test distribution.

7.0.2 DS due to population-based selection bias

What are the causes of such DS? We considered acquisition-related DS in the methods developed in this thesis. Much of the DS robustness literature in medical image analysis has also focused on such DS. However, several other types of DS are also pertinent in medical imaging, even within the covariate shift umbrella (that is, shifts in the input distribution) (see Sec. 1.3.2). One of these shifts is due to population-based selection bias in the training data. Common causes for such bias can be due to factors such as age (training distribution consists of images of adults, but the test image is of an infant or an elderly person), sex, ethnicity, among others.

Difficulty in avoiding such DS: Such DS are also likely to cause performance degradation in CNN-based analysis methods [245]↑. As well, preventing such selection bias in the training dataset could be very difficult. That is, it is plausible that collection of a large enough training dataset that encompasses all population demographics may be infeasible.

How to drive TTA to tackle such DS?: We believe that the TTA setting could be ideally suited to improve model performance when faced with DS caused by selection biases in the training data. Let us consider two questions that must be answered to implement TTA in this setting. The first question is how to drive adaptation at test time, without access to a labelled dataset from the test distribution. The question can likely be answered using similar helper models as in chapters 5, 6 or as proposed in other works in the literature [68]↑, [123] ↑. However, this hypothesis is yet to be validated in the literature, to the best of our knowledge.

Design of the normalization sub-network: The second question is which parameters to adapt for TTA to tackle population-based DS. To answer this, we note that the role of the normalization sub-network, N_ϕ, in our TTA approaches (chapters 5, 6) was to map the given test image to a normalized image that is similar to the normalized versions of training images. Once such a mapping is achieved, the rest of the task CNN could map the normalized image to the correct prediction. For acquisition-related DS, N_ϕ was designed to be a shallow CNN with a small receptive field, and outputted an image with the same dimensionality as the input image. This design choice enabled N_ϕ to flexibly model contrast transformations without introducing substantial structural changes in the image. The design of this sub-network would have to be modified to tackle population-related DS. For tackling DS due to a selection bias by age, for instance, N_ϕ can potentially be designed to output a scale factor or a diffeomorphic deformation field [238]↑to model age-related changes in anatomical shapes. Furthermore, some DS may require adaptation in deeper layers to account for effects of demographic variables on the semantic information a CNN extracts for a given task. For DS for which it is unclear how to design N_ϕ, it may be interesting to adapt all the task CNN parameters at test time and disincentivize large deviations from the training distribution optimal parameters [244]↑, [124]↑.

A special example of population-based DS: Another scenario for population-based DS could be when the training dataset consists of images of healthy individuals, while the test image is from a diseased subject. In a practical setting, such information may be unknown during image acquisition or automated analysis. We discuss this scenario separately in Sec. 7.0.5. In contrast, in this section, we considered population-based DS that are known to exist at test time, and discussed the potential of TTA approaches to improve robustness of analysis methods in their presence.

7.0.3 DS due to task-specific selection biases

What are the causes of such DS? A different type of selection bias than the one discussed in Sec. 7.0.2 can occur in the case of image enhancement tasks, such as image reconstruction from undersampled measurements, image super-resolution, image denoising, among others. As described in Sec. 1.1.1, the goal in such tasks is to obtain a mapping from corrupted images to enhanced, corruption-free images. For such tasks, the training dataset is typically constructed in a self-supervised manner. That is, we have access to a set of enhanced images, and corresponding corrupted images are generated by following a known corruption process. Once the model is trained, it is fed with new corrupted images and predicts the corresponding unknown enhanced images. A DS in the model's input distribution can occur if the self-supervised process generating the model's training dataset does not adequately represent the corruptions observed at test time. Thus, the task-specific DS in image enhancement problems are shifts in the input image's corruption distribution. Furthermore, for some tasks, it may be feasible to acquire paired corrupted and enhanced images to form the training dataset. In such cases, depending on the variability covered in the training dataset, the likelihood of observing task-specific DS at test time may be even higher.

Examples of task-specific DS: In MRI reconstruction from undersampled k-space measurements, training and test measurements may differ in terms of the type of undersampling masks [246]↑, [247]↑. For image super-resolution, training and test sets may differ in the type of undersampling kernels [128]↑, [248]↑. For image denoising, DS may arise due to difference in type of noise in the input images [249]↑.

Difficulty in avoiding such DS: In some of these examples (e.g. MRI reconstruction), the task specific factors of variations (e.g. undersampling mask) can be classified as related to the acquisition protocol. Such factors have to be treated differently than the generic acquisition-related DS that manifest primarily as contrast changes and that were considered in the methods developed in this thesis. One solution to deal with task-specific DS may be carefully design self-supervised training datasets such that all plausible corruptions are included. However, due to the high variability in medical imaging acquisition protocols, it is difficult to completely rule out the possibility of encountering new types of corruptions at test time.

TTA literature to tackle task-specific DS: One TTA approach is to transform the test image with different corruption patterns to the corresponding corrupted image with training-like corruptions. Following this approach, for image super-resolution, cycle-consistency-based estimation of a correction filter has been proposed to transform low-resolution (LR) test images to resemble LR images seen during training [128]↑.

Another TTA approach is to forgo reliance on training corruption patterns as well as on paired clean-corrupted images in the test distribution. In these works, learning takes place directly on a set of corrupted images. In the denoising literature, for instance, it has been proposed to achieve this by carrying out unsupervised learning using noisy images only, without depending on corresponding clean images. Specifically, in the Gaussian noise setting, [129]↑estimate the distance between a given noisy and unknown clean image, via Stein's unbiased risk estimator [250]↑. The denoising CNN is trained with this estimated loss. As such learning does not depend on noisy and clean image pairs, it can be done either directly for the test images or can be used to drive TTA of a learned model to resolve performance degradation due to lack of DS robustness.

Yet another approach that forgoes reliance on paired clean-corrupted images is that of Bayesian image enhancement [251]↑, [157]↑. Here, a model of clean images is learned, and corrupted images are transformed into images that are clean according to the learned model. Owing to the absence of particular types of corrupted images during training, these methods work well across a wide spectrum of corruption patterns in the test images. However, such methods may be susceptible to other types of DS, as elaborated in Sec. 7.0.6.

7.0.4 Acquisition-related DS in image enhancement problems

The causes for task-specific DS discussed in Sec. 7.0.3 are additional DS sources for image enhancement tasks, along with the acquisition-related and population-related DS discussed previously. Thus, even in cases where the training dataset of image enhancement models covers corruptions types observed at test time, the test images may still be from a shifted distribution in that they may be acquired from a different scanner or using different acquisition protocol parameters.

TTA literature to tackle acquisition-related DS in image enhancement problems: As discussed before, acquisition-related DS has been extensively studied for several analysis tasks. The same is the case with image enhancement tasks. Here, one common idea is to employ task-specific losses to drive TTA to improve model performance when faced with DS. An example of such task-specific losses is k-space data consistency in MRI reconstruction CNNs [125]↑, [126]↑. For image super-resolution, [252]↑, [253]↑leverage the fact that medical images are often acquired as 3D volumes, with high in-plane resolution and low through-plane resolution, to generate 'low-resolution'-'high-resolution' training pairs directly from the test image at hand.

Can the normalization sub-network model only contrast changes?: For the task-specific TTA losses described above, either all task CNN parameters [125]↑, [126]↑, [252]↑or the parameters of a relatively deep feature extraction sub-network [253]↑are adapted at test time. This choice is unlike the shallow normalization module that is adapted for each test image in chapters 5, 6. This indicates that increased flexibility may be required at test time for achieving TTA in image enhancement tasks as compared to tasks such as segmentation, registration or classification that take enhanced images as inputs and extract information from them. For the latter group of tasks, the input images from training and test distributions differ in terms of their acquisition details; thus, a pixel-wise contrast transformation might suffice to map one to the other. On the other hand, in image enhancement tasks, such a pixel-wise contrast transformation exists between the training and test distributions in the output space, but may be insufficient to model the relationship between the corresponding corrupted input images. Achieving stable TTA despite the increased flexibility afforded at test time might potentially require introduction of regularization constraints [244]↑, [124]↑.

7.0.5 DS due to imaging artifacts or presence of disease

So far, we have DS due to causes that are known when an algorithm is asked to analyze a test image. Accordingly, this knowledge could be leveraged to dictate TTA design choices (design of adaptable module, TTA loss). A more challenging scenario is one where a DS occurs due to causes that cannot be predicted in advance. Examples of such DS are presence of imaging artifacts (for instance, due to patient motion during the image acquisition [254]↑) or presence of anatomical anomalies that were not present in the training dataset. In such cases, the algorithm should be able to detect such DS by itself, and flag the test image without making a prediction for the task. Previous work has considered supervised detection of motion artifacts [255]↑, [256]↑. However, such approaches may be unable to detect artifacts beyond those present during training. We believe that the out-of-distribution (OOD) detection setting (Sec. 2.3.1) is most relevant for such DS, while TTA is the most relevant setting for DS that are already known to exist at prediction time.

7.0.6 DS robustness of density estimation models

We have considered the DS robustness problem in model trained via supervised learning. CNN-based density estimation models are also frequently employed in medical image analysis methods. Specifically, they play a central role in medical image enhancement tasks solved in a Bayesian framework, where they are used to model the probability density of enhanced images. In this setting, both implicit (e.g. generative adversarial networks [257]↑) as well as explicit (e.g. variational autoencoders [258]↑) density estimation models can be used. In this section, we discuss the DS robustness of such models, and ways to improve the same.

TTA of implicit density estimation models: In the former set of methods, the latent code of the generator is optimized [259]↑, [260]↑to provide an image that, when corrupted, is similar to the given corrupted image. This approach may potentially suffer from the representation error issue - that is, even for images from the training distribution, there exists a discrepancy between the optimized and the true enhanced image [259]↑. To remedy this, TTA of the generator weights along with the latent code has been suggested [134]↑, [261]↑. Taking this to the extreme and showcasing an inherent model prior in neural networks, TTA of completely unlearned networks has also been proposed [135]↑, [262]↑.

TTA of explicit density estimation models: Explicit density estimation models allow for iterative optimization directly in the image space [251]↑, [157]↑. However, such models are unreliable when faced with acquisition-related DS - that is, when the corrupted (undersampled / low-resolution, etc.) image is acquired with different acquisition protocol parameters or from a different scanner as compared to the enhanced images used for training the prior. TTA could be useful in such settings, but is unexplored in the literature to the best of our knowledge. Given access to enhanced images from the test distribution, TTA can be achieved by maximizing their (approximate) log-probability [263]↑, [264]↑. Similar to chapters 5, 6, a shallow sub-network could be adapted for modeling the intensity transformation between the samples from the training and test distributions.

TTA of density estimation models with access to only corrupted test images: The more challenging setting is if TTA has to be done when only corrupted versions of images from the test distribution are available. In this setting, it could be interesting to incorporate ideas from the literature on learning from corrupted data only [265]↑, [266]↑, [267]↑, [268]↑within the TTA framework.

An inherent DS in density estimation models: Additionally, we note that explicit density estimation models used for Bayesian image enhancement suffer from an inherent DS issue, that exists even in the absence of any acquisition-related DS. This is caused by the fact that such models are often trained only using samples from the true distribution of enhanced images, without any information regarding the corruptions to be corrected at test time. On one hand, such models are general and can be used to remove different types of corruptions. On the other hand, their behaviour at points that do not belong to the true distribution may be unreliable [164]↑. To remedy this, it could be helpful to introduce knowledge of expected corruptions while training the prior models [225]↑, [269]↑. Usage of samples from simple corrupted distributions (such as Gaussian noise) for training density models has been suggested [270]↑, [271]↑. As well, unrolled optimization methods [272]↑implicitly learn the gradient of log-prior-density at the corruptions observed during training, but are restricted only to such corruptions. Supervised density estimation [225]↑, [269]↑could potentially combine the accuracy of supervised methods with the generality of unsupervised ones.

Bibliography

[1] Nadine Barrie Smith and Andrew Webb, Introduction to Medical Imaging: Physics, Engineering and Clinical Applications, Cambridge Texts in Biomedical Engineering. Cambridge University Press, 2010.

[2] OECD INDICATORS, "Health at a glance 2017," 2017.

[3] A Van der Gijp, MF Van der Schaaf, IC Van der Schaaf, JCBM Huige, CJ Ravesloot, JPJ Van Schaik, and Th J Ten Cate, "Interpretation of radiological images: towards a framework of knowledge and skills," Advances in Health Sciences Education, vol. 19, no. 4, pp. 565–580, 2014.

[4] Tirath Y Patel and McKinley Glover, "The time is now: Revisiting the case for the 3-year radiology residency," Journal of the American College of Radiology, vol. 12, no. 5, pp. 481–483, 2015.

[5] Oliver Ruprecht, Philipp Weisser, Boris Bodelle, Hanns Ackermann, and Thomas J Vogl, "Mri of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy," European journal of radiology, vol. 81, no. 3, pp. 456–460, 2012.

[6] Anton S Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J Muehlematter, Andreas M Hötker, Ender Konukoglu, and Olivio F Donati, "Variability of manual segmentation of the prostate in axial t2-weighted mri: A multi-reader study," European journal of radiology, vol. 121, pp. 108716, 2019.

[7] Elizabeth A Krupinski, "Current perspectives in medical image perception," Attention, Perception, & Psychophysics, vol. 72, no. 5, pp. 1205–1217, 2010.

[8] Daniel J Mollura, Ezana M Azene, Anna Starikovsky, Aduke Thelwell, Sarah Iosifescu, Cary Kimble, Ann Polin, Brian S Garra, Kristen K DeStigter, Brad Short, et al., "White paper report of the rad-aid conference on international radiology for developing countries: identifying challenges, opportunities, and strategies for imaging services in the developing world," Journal of the American College of Radiology, vol. 7, no. 7, pp. 495–500, 2010.

[9] Guy Frija, Ivana Blažić, Donald P Frush, Monika Hierath, Michael Kawooya, Lluis Donoso-Bach, and Boris Brkljačić, "How to improve access to medical imaging in low-and middle-income countries?," EClinicalMedicine, vol. 38, pp. 101034, 2021.

[10] Kokou Adambounou, Victor Adjenou, Alex P Salam, Fabien Farin, Koffi Gilbert N'Dakena, Messanvi Gbeassor, and Philippe Arbeille, "A low-cost tele-imaging platform for developing countries," Frontiers in public health, vol. 2, pp. 135, 2014.

[11] Thierry Pun, Guido Gerig, and Osman Ratib, "Image analysis and computer vision in medicine," Computerized Medical Imaging and Graphics, vol. 18, no. 2, pp. 85–96, 1994.

[12] James S Duncan and Nicholas Ayache, "Medical image analysis: Progress over two decades and the challenges ahead," IEEE transactions on pattern analysis and machine intelligence, vol. 22, no. 1, pp. 85–106, 2000.

[13] Dzung L Pham, Chenyang Xu, and Jerry L Prince, "Current methods in medical image segmentation," Annual review of biomedical engineering, vol. 2, no. 1, pp. 315–337, 2000.

[14] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," Journal of digital imaging, vol. 32, no. 4, pp. 582–596, 2019.

[15] Tao Lei, Risheng Wang, Yong Wan, Bingtao Zhang, Hongying Meng, and Asoke K Nandi, "Medical image segmentation using deep learning: a survey," arXiv preprint arXiv:2009.13120, 2020.

[16] Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes, "Medical image registration," Physics in medicine & biology, vol. 46, no. 3, pp. R1, 2001.

[17] Francisco PM Oliveira and Joao Manuel RS Tavares, "Medical image registration: a review," Computer methods in biomechanics and biomedical engineering, vol. 17, no. 2, pp. 73–93, 2014.

[18] Grant Haskins, Uwe Kruger, and Pingkun Yan, "Deep learning in medical image registration: a survey," Machine Vision and Applications, vol. 31, no. 1, pp. 1–18, 2020.

[19] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen, "Medical image classification with convolutional neural network," in 2014 13th international conference on control automation robotics & vision (ICARCV). IEEE, 2014, pp. 844–848.

[20] Gengsheng Lawrence Zeng, Medical image reconstruction: a conceptual tutorial., Springer, 2010.

[21] Hai-Miao Zhang and Bin Dong, "A review on deep learning in medical image reconstruction," Journal of the Operations Research Society of China, pp. 1–30, 2020.

[22] Jithin Saji Isaac and Ramesh Kulkarni, "Super resolution techniques for medical image processing," in 2015 International Conference on Technologies for Sustainable Development (ICTSD). IEEE, 2015, pp. 1–6.

[23] Reza Amini Gougeh, Tohid Yousefi Rezaii, and Ali Farzamnia, "Medical image enhancement and deblurring," in Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019. Springer, 2021, pp. 543–554.

[24] Sameera V Mohd Sagheer and Sudhish N George, "A review on medical image denoising algorithms," Biomedical signal processing and control, vol. 61, pp. 102036, 2020.

[25] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang, "A review on medical imaging synthesis using deep learning and its clinical applications," Journal of Applied Clinical Medical Physics, vol. 22, no. 1, pp. 11–36, 2021.

[26] Ruth E Baker, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem, "Mechanistic models versus machine learning, a fight worth fighting for the biological community?," Biology letters, vol. 14, no. 5, pp. 20170660, 2018.

[27] Daniel Rueckert and Julia A Schnabel, "Model-based and data-driven strategies in medical image computing," Proceedings of the IEEE, vol. 108, no. 1, pp. 110–124, 2019.

[28] Hákon Gudbjartsson and Samuel Patz, "The rician distribution of noisy mri data," Magnetic resonance in medicine, vol. 34, no. 6, pp. 910–914, 1995.

[29] Qolamreza R Razlighi, Nasser Kehtarnavaz, and Siamak Yousefi, "Evaluating similarity measures for brain image registration," Journal of visual communication and image representation, vol. 24, no. 7, pp. 977–987, 2013.

[30] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, and Paul Suetens, "Automated model-based tissue classification of mr images of the brain," IEEE transactions on medical imaging, vol. 18, no. 10, pp. 897–908, 1999.

[31] Michael Lustig, David Donoho, and John M Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 58, no. 6, pp. 1182–1195, 2007.

[32] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache, “Diffeomorphic demons: Efficient non-parametric image registration,” NeuroImage, vol. 45, no. 1, pp. S61–S72, 2009.

[33] Y. Zhang, M. Brady, and S. Smith, “Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm,” IEEE Transactions on Medical Imaging, vol. 20, no. 1, pp. 45–57, 2001.

[34] V. Vapnik, “Principles of risk minimization for learning theory,” in Advances in neural information processing systems, 1992.

[35] Ulrike Von Luxburg and Bernhard Schölkopf, “Statistical learning theory: Models, concepts, and results,” in Handbook of the History of Logic, vol. 10, pp. 651–706. Elsevier, 2011.

[36] Kunihiko Fukushima and Sei Miyake, “Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition,” in Competition and cooperation in neural nets, pp. 267–285. Springer, 1982.

[37] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio, “Object recognition with gradient-based learning,” in Shape, contour and grouping in computer vision, pp. 319–345. Springer, 1999.

[38] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, “Deep sparse rectifier neural networks,” in Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[39] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[40] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International conference on machine learning. PMLR, 2015.

[41] Vivienne Sze, Yu-Hsin Chen, Joel Emer, Amr Suleiman, and Zhengdong Zhang, “Hardware for machine learning: Challenges and opportunities,” in 2017 IEEE Custom Integrated Circuits Conference (CICC). IEEE, 2017, pp. 1–8.

[42] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al., “Evaluation of prostate segmentation algorithms for mri: the promise12 challenge,” Medical image analysis, 2014.

[43] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al., “Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?,” IEEE transactions on medical imaging, vol. 37, no. 11, pp. 2514–2525, 2018.

[44] Guoyan Zheng, Chengwen Chu, Daniel L Belavỳ, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, et al., “Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge,” Medical image analysis, vol. 35, pp. 327–344, 2017.

[45] Dinggang Shen, Guorong Wu, and Heung-Il Suk, “Deep learning in medical image analysis,” Annual review of biomedical engineering, vol. 19, pp. 221–248, 2017.

[46] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, “A survey on deep learning in medical image analysis,” Medical image analysis, vol. 42, pp. 60–88, 2017.

[47] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers, “A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,” Proceedings of the IEEE, 2021.

[48] Ozan Oktay, Jay Nanavati, Anton Schwaighofer, David Carter, Melissa Bristow, Ryutaro Tanno, Rajesh Jena, Gill Barnett, David Noble, Yvonne Rimmer, et al., "Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers," JAMA network open, vol. 3, no. 11, pp. e2027426–e2027426, 2020.

[49] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó, "The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database," NPJ digital medicine, vol. 3, no. 1, pp. 1–8, 2020.

[50] Mohammad Hosein Rezazade Mehrizi, Peter van Ooijen, and Milou Homan, "Applications of artificial intelligence (ai) in diagnostic radiology: a technography study," European radiology, vol. 31, no. 4, pp. 1805–1811, 2021.

[51] Kicky G van Leeuwen, Steven Schalekamp, Matthieu JCM Rutten, Bram van Ginneken, and Maarten de Rooij, "Artificial intelligence in radiology: 100 commercially available products and their scientific evidence," European radiology, vol. 31, no. 6, pp. 3797–3804, 2021.

[52] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer, Dataset shift in machine learning, Mit Press, 2009.

[53] D. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," Nature Communications, vol. 11, no. 1, pp. 1–10, 2020.

[54] G Preboske et al., "Common MRI acquisition non-idealities significantly impact the output of the boundary shift integral method of measuring brain atrophy on serial MRI," Neuroimage, vol. 30, no. 4, pp. 1196–1202, 2006.

[55] J Jovicich et al., "Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data," Neuroimage, vol. 30, no. 2, pp. 436–443, 2006.

[56] Ben Glocker, Robert Robinson, Daniel C Castro, Qi Dou, and Ender Konukoglu, "Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects," arXiv preprint arXiv:1910.04597, 2019.

[57] Jürgen Weese and Cristian Lorenz, "Four challenges in medical image analysis from an industrial perspective," 2016.

[58] Ehab A AlBadawy, Ashirbani Saha, and Maciej A Mazurowski, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," Medical physics, vol. 45, no. 3, pp. 1150–1158, 2018.

[59] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, Milica G Kramberger, et al., "The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study," Medical Image Analysis, vol. 66, pp. 101714, 2020.

[60] W. Yan, L. Huang, L. Xia, S. Gu, F. Yan, Y. Wang, and Q. Tao, "Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners," Radiology: Artificial Intelligence, vol. 2, no. 4, pp. e190195, 2020.

[61] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study," PLoS medicine, vol. 15, no. 11, pp. e1002683, 2018.

[62] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi, "An empirical framework for domain generalization in clinical settings," in Proceedings of the Conference on Health, Inference, and Learning, 2021.

[63] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proceedings of the ACM conference on health, inference, and learning, 2020, pp. 151–159.

[64] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," Proceedings of the National Academy of Sciences, vol. 117, no. 23, pp. 12592–12594, 2020.

[65] Paul T Fillmore, Michelle C Phillips-Meek, and John E Richards, "Age-specific mri brain and head templates for healthy adults from 20 through 89 years of age," Frontiers in aging neuroscience, vol. 7, pp. 44, 2015.

[66] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain mr segmentation across scanners and protocols," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018.

[67] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," Medical Image Analysis, vol. 68, pp. 101907, 2021.

[68] Y. He, A. Carass, L. Zuo, B.E. Dewey, and J. L. Prince, "Autoencoder based self-supervised test-time adaptation for medical image analysis," Medical Image Analysis, p. 102136, 2021.

[69] Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al., "Results of the 2020 fastmri challenge for machine learning mr image reconstruction," IEEE transactions on medical imaging, vol. 40, no. 9, pp. 2306–2317, 2021.

[70] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," IEEE transactions on medical imaging, vol. 34, no. 10, pp. 1993–2024, 2014.

[71] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," Scientific data, vol. 4, no. 1, pp. 1–13, 2017.

[72] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al., "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," arXiv preprint arXiv:1811.02629, 2018.

[73] A. Van Opbroek, A. Ikram, M. Vernooij, and M. De Bruijne, "Transfer learning improves supervised image segmentation across imaging protocols," IEEE transactions on medical imaging, vol. 34, no. 5, pp. 1018–1030, 2014.

[74] N. Tajbakhsh, J. Shin, S. Gurudu, T. Hurst, C. B. Kendall, M. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[75] Namgyu Ho and Yoon-Chul Kim, "Evaluation of transfer learning in deep convolutional neural network models for cardiac short axis slice classification," Scientific reports, vol. 11, no. 1, pp. 1–11, 2021.

[76] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris, "Spottune: transfer learning through adaptive fine-tuning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4805–4814.

[77] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[78] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio, "Transfusion: Understanding transfer learning for medical imaging," arXiv preprint arXiv:1902.07208, 2019.

[79] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[80] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu, “Contrastive learning of global and local features for medical image segmentation with limited annotations,” arXiv preprint arXiv:2006.10511, 2020.

[81] Sebastian Thrun, “Lifelong learning algorithms,” in Learning to learn, pp. 181–209. Springer, 1998.

[82] Robert M French, “Catastrophic forgetting in connectionist networks,” Trends in cognitive sciences, vol. 3, no. 4, pp. 128–135, 1999.

[83] Anthony Robins, “Catastrophic forgetting, rehearsal and pseudorehearsal,” Connection Science, vol. 7, no. 2, pp. 123–146, 1995.

[84] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim, “Continual learning with deep generative replay,” arXiv preprint arXiv:1705.08690, 2017.

[85] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al., “Overcoming catastrophic forgetting in neural networks,” Proceedings of the national academy of sciences, vol. 114, no. 13, pp. 3521–3526, 2017.

[86] Garrett Wilson and Diane J Cook, “A survey of unsupervised deep domain adaptation,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 5, pp. 1–46, 2020.

[87] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto, “Unified deep supervised domain adaptation and generalization,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5715–5725.

[88] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al., “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” in International conference on information processing in medical imaging. Springer, 2017.

[89] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, “Domain-adversarial training of neural networks,” The journal of machine learning research, vol. 17, no. 1, pp. 2096–2030, 2016.

[90] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, “Synsegnet: Synthetic segmentation without target modality ground truth,” IEEE transactions on medical imaging, vol. 38, no. 4, pp. 1016–1025, 2018.

[91] Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Ertunc Erdil, Anton Becker, Olivio Donati, and Ender Konukoglu, “Semi-supervised task-driven data augmentation for medical image segmentation,” Medical Image Analysis, vol. 68, pp. 101934, 2021.

[92] Yabin Zhang, Haojian Zhang, Bin Deng, Shuai Li, Kui Jia, and Lei Zhang, “Semi-supervised models are strong unsupervised domain adaptation learners,” arXiv preprint arXiv:2106.00417, 2021.

[93] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf, “Domain generalization via invariant feature representation,” in International Conference on Machine Learning. PMLR, 2013, pp. 10–18.

[94] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi, “Domain generalization for object recognition with multi-task autoencoders,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2551–2559.

[95] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot, “Domain generalization with adversarial feature learning,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5400–5409.

[96] M. W. Lafarge, J. Pluim, K. Eppenhof, and M. Veta, “Learning domain-invariant representations of histological images,” Frontiers in medicine, vol. 6, pp. 162, 2019.

[97] X. Liu, S. Thermos, A. O'Neil, and S. Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021.

[98] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," in International Conference on Learning Representations, 2018.

[99] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang, "Self-challenging improves cross-domain generalization," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 124–140.

[100] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert, "Causality-inspired single-source domain generalization for medical image segmentation," arXiv preprint arXiv:2111.12525, 2021.

[101] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu, and Z. Xu, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," IEEE Transactions on Medical Imaging, vol. 39, no. 7, pp. 2531–2540, 2020.

[102] Amod Jog, Andrew Hoopes, Douglas N Greve, Koen Van Leemput, and Bruce Fischl, "Psacnn: Pulse sequence adaptive fast whole brain segmentation," NeuroImage, vol. 199, pp. 553–569, 2019.

[103] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese, "Generalizing to unseen domains via adversarial data augmentation," arXiv preprint arXiv:1805.12018, 2018.

[104] Chen Chen, Chen Qin, Huaqi Qiu, Cheng Ouyang, Shuo Wang, Liang Chen, Giacomo Tarroni, Wenjia Bai, and Daniel Rueckert, "Realistic adversarial data augmentation for mr image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 667–677.

[105] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi, "Generalizing across domains via cross-gradient training," arXiv preprint arXiv:1804.10745, 2018.

[106] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer, "Robust and generalizable visual representation learning via random convolutions," in International Conference on Learning Representations, 2020.

[107] B. Billot, D. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, Adrian Dalca, and Juan Eugenio Iglesias, "Synthseg: Domain randomisation for segmentation of brain mri scans of any contrast and resolution," 2021.

[108] Bernhard Schölkopf, "Causality for machine learning," arXiv preprint arXiv:1911.10500, 2019.

[109] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré, "Selecting data augmentation for simulating interventions," in International Conference on Machine Learning. PMLR, 2021, pp. 4555–4562.

[110] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Learning to generalize: Meta-learning for domain generalization," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[111] Q. Dou, D. C. de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in Advances in Neural Information Processing Systems, 2019.

[112] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao, "Generalizable semantic segmentation via model-agnostic learning and target-specific normalization," arXiv preprint arXiv:2003.12296, vol. 2, no. 3, pp. 6, 2020.

[113] Taco Cohen and Max Welling, "Group equivariant convolutional networks," in International conference on machine learning. PMLR, 2016, pp. 2990–2999.

[114] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow, "Harmonic networks: Deep translation and rotation equivariance," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5028–5037.

[115] Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis, "Polar transformer networks," in International Conference on Learning Representations, 2018.

[116] Kai Sheng Tai, Peter Bailis, and Gregory Valiant, "Equivariant transformer networks," in International Conference on Machine Learning. PMLR, 2019, pp. 6086–6095.

[117] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2020.

[118] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[119] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie, "Are transformers more robust than cnns?," Advances in Neural Information Processing Systems, vol. 34, 2021.

[120] Patrice Simard, Bernard Victorri, Yann Le Cun, and John Denker, "Tangent prop: a formalism for specifying selected invariances in an adaptive network," in Proceedings of the 4th International Conference on Neural Information Processing Systems, 1991, pp. 895–903.

[121] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in Icml, 2011.

[122] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in International Conference on Machine Learning. PMLR, 2020.

[123] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in International Conference on Learning Representations, 2021.

[124] Aurick Zhou and Sergey Levine, "Bayesian adaptation for covariate shift," Advances in Neural Information Processing Systems, vol. 34, 2021.

[125] J. Zhang, Z. Liu, S. Zhang, H. Zhang, P. Spincemaille, T. Nguyen, M. Sabuncu, and Y. Wang, "Fidelity imposed network edit (fine) for solving ill-posed image reconstruction," NeuroImage, vol. 211, pp. 116579, 2020.

[126] D. Gilton, G. Ongie, and R. Willett, "Model adaptation for inverse problems in imaging," IEEE Transactions on Computational Imaging, vol. 7, pp. 661–674, 2021.

[127] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, et al., "Interactive medical image segmentation using deep learning with image-specific fine tuning," IEEE transactions on medical imaging, vol. 37, no. 7, pp. 1562–1573, 2018.

[128] S. A.Hussein, T. Tirer, and R. Giryes, "Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[129] Shakarim Soltanayev and Se Young Chun, "Training deep learning based denoisers without ground truth data," in Advances in Neural Information Processing Systems, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, Curran Associates, Inc.

[130] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn, "Adaptive risk minimization: Learning to adapt to domain shift," Advances in Neural Information Processing Systems, vol. 34, 2021.

[131] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," Pattern Recognition, vol. 80, pp. 109–117, 2018.

[132] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," Advances in Neural Information Processing Systems, 2020.

[133] C. Burns and J. Steinhardt, “Limitations of post-hoc feature alignment for robustness,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[134] S. A. Hussein, T. Tirer, and R. Giryes, “Image-adaptive gan based reconstruction,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[135] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.

[136] J. Yoo, K. H. Jin, H. Gupta, J. Yerly, M. Stuber, and M. Unser, “Time-dependent deep image prior for dynamic mri,” IEEE Transactions on Medical Imaging, 2021.

[137] V. Jain and E. Learned-Miller, “Online domain adaptation of a pre-trained cascade of classifiers,” in CVPR 2011. IEEE, 2011.

[138] B. Chidlovskii, S. Clinchant, and G. Csurka, “Domain adaptation in the absence of source domain data,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[139] J. Liang, D. Hu, and J. Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in International Conference on Machine Learning. PMLR, 2020.

[140] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed, “Source-relaxed domain adaptation for image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020.

[141] Philipp Krähenbühl and Vladlen Koltun, “Efficient inference in fully connected crfs with gaussian edge potentials,” Advances in neural information processing systems, vol. 24, pp. 109–117, 2011.

[142] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker, “Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation,” Medical image analysis, vol. 36, pp. 61–78, 2017.

[143] Agostina J Larrazabal, César Martínez, Ben Glocker, and Enzo Ferrante, “Post-dae: Anatomically plausible segmentation via post-processing with denoising autoencoders,” IEEE Transactions on Medical Imaging, vol. 39, no. 12, pp. 3813–3820, 2020.

[144] Melih Engin, Robin Lange, Andras Nemes, Sadaf Monajemi, Milad Mohammadzadeh, Chin Kong Goh, Tian Ming Tu, Benjamin YQ Tan, Prakash Paliwal, Leonard LL Yeo, et al., “Agan: An anatomy corrector conditional generative adversarial network,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 708–717.

[145] Y Zhang and M Brady, “smith sm. segmentation of brain mr images through a hidden markov random field model and the expectation maximization algorithm,” IEEE Trans Med Imaging, vol. 20, no. 1, pp. 45–57, 2001.

[146] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al., “Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain,” Neuron, vol. 33, no. 3, pp. 341–355, 2002.

[147] Koen Van Leemput, Frederik Maes, Dirk Vandermeulen, Alan Colchester, and Paul Suetens, “Automated segmentation of multiple sclerosis lesions by model outlier detection,” IEEE transactions on medical imaging, vol. 20, no. 8, pp. 677–688, 2001.

[148] Suyash P Awate, Tolga Tasdizen, Norman Foster, and Ross T Whitaker, “Adaptive markov modeling for mutual-information-based, unsupervised mri brain-tissue classification,” Medical Image Analysis, vol. 10, no. 5, pp. 726–739, 2006.

[149] Mert R Sabuncu, BT Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland, “A generative model for image segmentation based on label fusion,” IEEE transactions on medical imaging, vol. 29, no. 10, pp. 1714–1729, 2010.

[150] Oula Puonti, Juan Eugenio Iglesias, and Koen Van Leemput, "Fast and sequence-adaptive whole-brain segmentation using parametric bayesian modeling," NeuroImage, vol. 143, pp. 235–249, 2016.

[151] Adrian V Dalca, Evan Yu, Polina Golland, Bruce Fischl, Mert R Sabuncu, and Juan Eugenio Iglesias, "Unsupervised deep learning for bayesian brain mri segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 356–365.

[152] Mikael Agn, Per Munck af Rosenschöld, Oula Puonti, Michael J Lundemann, Laura Mancini, Anastasia Papadaki, Steffi Thust, John Ashburner, Ian Law, and Koen Van Leemput, "A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning," Medical image analysis, vol. 54, pp. 220–237, 2019.

[153] Mikael Brudfors, Yaël Balbastre, and John Ashburner, "Nonlinear markov random fields learned via backpropagation," in International Conference on Information Processing in Medical Imaging. Springer, 2019, pp. 805–817.

[154] Adrian V Dalca, John Guttag, and Mert R Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9290–9299.

[155] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu, "Generalized out-of-distribution detection: A survey," arXiv preprint arXiv:2110.11334, 2021.

[156] Xiaoran Chen and Ender Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," arXiv preprint arXiv:1806.04972, 2018.

[157] Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," Medical image analysis, vol. 64, pp. 101713, 2020.

[158] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," Advances in neural information processing systems, vol. 31, 2018.

[159] Ev Zisselman and Aviv Tamar, "Deep residual flow for out of distribution detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13994–14003.

[160] Ivan Kobyzev, Simon Prince, and Marcus Brubaker, "Normalizing flows: An introduction and review of current methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[161] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang, "Deep structured energy based models for anomaly detection," in International Conference on Machine Learning. PMLR, 2016, pp. 1100–1109.

[162] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li, "Energy-based out-of-distribution detection," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 21464–21475, Curran Associates, Inc.

[163] Hyunsun Choi, Eric Jang, and Alexander A Alemi, "Waic, but why? generative ensembles for robust anomaly detection," arXiv preprint arXiv:1810.01392, 2018.

[164] E. Nalisnick, A. Matsukawa, Y. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?," in International Conference on Learning Representations, 2019.

[165] W. Morningstar, C. Ham, A. Gallagher, B. Lakshminarayanan, A. Alemi, and J. Dillon, "Density of states estimation for out of distribution detection," in International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

[166] E. Erdil, K. Chaitanya, N. Karani, and E. Konukoglu, "Task-agnostic out-of-distribution detection using kernel density estimation," in Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis, Cham, 2021, Springer International Publishing.

[167] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song, “Using self-supervised learning can improve model robustness and uncertainty,” arXiv preprint arXiv:1906.12340, 2019.

[168] Camila Gonzalez and Anirban Mukhopadhyay, “Self-supervised out-of-distribution detection for cardiac cmr segmentation,” in Medical Imaging with Deep Learning, 2021.

[169] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, “Deep anomaly detection with outlier exposure,” arXiv preprint arXiv:1812.04606, 2018.

[170] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin, “Training confidence-calibrated classifiers for detecting out-of-distribution samples,” arXiv preprint arXiv:1711.09325, 2017.

[171] Shiyu Liang, Yixuan Li, and R Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” in International Conference on Learning Representations, 2018.

[172] Hui Zhang, Sharath Cholleti, Sally A Goldman, and Jason E Fritts, “Meta-evaluation of image segmentation using machine learning,” in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06). IEEE, 2006, vol. 1, pp. 1138–1145.

[173] Timo Kohlberger, Vivek Singh, Chris Alvino, Claus Bahlmann, and Leo Grady, “Evaluating segmentation error without ground truth,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2012, pp. 528–536.

[174] Chao Huang, Qingbo Wu, and Fanman Meng, “Qualitynet: Segmentation quality evaluation with deep convolutional networks,” in 2016 Visual Communications and Image Processing (VCIP). IEEE, 2016, pp. 1–4.

[175] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker, “Reverse classification accuracy: predicting segmentation performance in the absence of ground truth,” IEEE transactions on medical imaging, vol. 36, no. 8, pp. 1597–1606, 2017.

[176] Terrance DeVries and Graham W Taylor, “Leveraging uncertainty estimates for predicting segmentation quality,” arXiv preprint arXiv:1807.00502, 2018.

[177] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer’s Disease Neuroimaging Initiative, et al., “Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control,” NeuroImage, vol. 195, pp. 11–22, 2019.

[178] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay, “Adversarial attacks and defences: A survey,” arXiv preprint arXiv:1810.00069, 2018.

[179] Kyriakos D Apostolidis and George A Papakostas, “A survey on adversarial deep learning robustness in medical image analysis,” Electronics, vol. 10, no. 17, pp. 2132, 2021.

[180] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam, “Adversarial attacks against medical deep learning systems,” arXiv preprint arXiv:1804.05296, 2018.

[181] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani, “Nci-isbi 2013 challenge: Automated segmentation of prostate structures.,” 2015.

[182] A. S. Becker, A. Cornelius, C. S. Reiner, D. Stocker, E. J. Ulbrich, B. K. Barth, A. Mortezavi, D. Eberli, and O. F. Donati, “Direct comparison of pi-rads version 2 and version 1 regarding interreader agreement and diagnostic accuracy for the detection of clinically significant prostate cancer,” European journal of radiology, vol. 94, pp. 58–63, 2017.

[183] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, et al., “Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge,” IEEE Transactions on Medical Imaging, 2021.

[184] Caroline Petitjean, Maria A Zuluaga, Wenjia Bai, Jean-Nicolas Dacher, Damien Grosgeorge, Jérôme Caudron, Su Ruan, Ismail Ben Ayed, M Jorge Cardoso, Hsiang-Chou Chen, et al., “Right ventricle segmentation from cardiac mri: a collation study,” Medical image analysis, vol. 19, no. 1, pp. 187–202, 2015.

[185] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener, et al., "Spinal cord grey matter segmentation challenge," Neuroimage, vol. 152, pp. 312–329, 2017.

[186] D. Van Essen, S.Smith, D. Barch, T. Behrens, E. Yacoub, K. Ugurbil, Wu-Minn HCP Consortium, et al., "The wu-minn human connectome project: an overview," Neuroimage, vol. 80, pp. 62–79, 2013.

[187] A. Di Martino, C. G. Yan, Q. Li, E. Denio, F. Castellanos, K. Alaerts, J. Anderson, M. Assaf, S. Bookheimer, M. Dapretto, et al., "The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism," Molecular psychiatry, vol. 19, no. 6, pp. 659–667, 2014.

[188] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett, "The alzheimer's disease neuroimaging initiative," Neuroimaging Clinics of North America, vol. 15, no. 4, pp. 869, 2005.

[189] H. Kuijf, M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, J. Cardoso, A. Casamitjana, et al., "Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge," IEEE transactions on medical imaging, vol. 38, no. 11, pp. 2556–2568, 2019.

[190] Josephin Gawlitza, Martin Reiss-Zimmermann, Gregor Thörmer, Alexander Schaudinn, Nicolas Linder, Nikita Garnov, Lars-Christian Horn, Do Hoang Minh, Roman Ganzer, Jens-Uwe Stolzenburg, et al., "Impact of the use of an endorectal coil for 3 t prostate mri on image quality and cancer detection rate," Scientific reports, vol. 7, no. 1, pp. 1–8, 2017.

[191] Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J Fütterer, "Esur prostate mr guidelines 2012," European radiology, vol. 22, no. 4, pp. 746–757, 2012.

[192] Mark A Griswold, Peter M Jakob, Robin M Heidemann, Mathias Nittka, Vladimir Jellus, Jianmin Wang, Berthold Kiefer, and Axel Haase, "Generalized autocalibrating partially parallel acquisitions (grappa)," Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 47, no. 6, pp. 1202–1210, 2002.

[193] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger, "Sense: sensitivity encoding for fast mri," Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 42, no. 5, pp. 952–962, 1999.

[194] B. Fischl, "Freesurfer," Neuroimage, vol. 62, no. 2, pp. 774–781, 2012.

[195] Joanna M Wardlaw, Eric E Smith, Geert J Biessels, Charlotte Cordonnier, Franz Fazekas, Richard Frayne, Richard I Lindley, John T O'Brien, Frederik Barkhof, Oscar R Benavente, et al., "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," The Lancet Neurology, vol. 12, no. 8, pp. 822–838, 2013.

[196] D. Marcus, T. Wang, J. Parker, J. Csernansky, J. Morris, and R. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," Journal of Cognitive Neuroscience, vol. 19, no. 9, pp. 1498–1507, 2007.

[197] V. Fonov, A. Evans, K. Botteron, R. Almli, R. McKinstry, L. Collins, Brain Development Cooperative Group, et al., "Unbiased average age-appropriate atlases for pediatric studies," Neuroimage, vol. 54, no. 1, pp. 313–327, 2011.

[198] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou, "Revisiting batch normalization for practical domain adaptation," arXiv preprint arXiv:1603.04779, 2016.

[199] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo, "Autodial: Automatic domain alignment layers," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5067–5075.

[200] Hakan Bilen and Andrea Vedaldi, "Universal representations: The missing link between faces, text, planktons, and cat breeds," arXiv preprint arXiv:1701.07275, 2017.

[201] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi, "Learning multiple visual domains with residual adapters," arXiv preprint arXiv:1705.08045, 2017.

[202] Ying Zhuge and Jayaram K Udupa, "Intensity standardization simplifies brain mr image segmentation," Computer vision and image understanding, vol. 113, no. 10, pp. 1095–1103, 2009.

[203] Neil L Weisenfeld and SK Warfteld, "Normalization of joint image-intensity statistics in mri using the kullback-leibler divergence," in 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821). IEEE, 2004, pp. 101–104.

[204] Xiao Han and Bruce Fischl, "Atlas renormalization for improved brain mr image segmentation across scanner platforms," IEEE transactions on medical imaging, vol. 26, no. 4, pp. 479–486, 2007.

[205] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015.

[206] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[207] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[208] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," Distill, vol. 1, no. 10, pp. e3, 2016.

[209] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 fourth international conference on 3D vision (3DV). IEEE, 2016.

[210] László G Nyúl, Jayaram K Udupa, and Xuan Zhang, "New variants of a method of mri scale standardization," IEEE transactions on medical imaging, vol. 19, no. 2, pp. 143–150, 2000.

[211] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," Journal of machine learning research, vol. 11, no. 12, 2010.

[212] Siavash Arjomand Bigdeli, Matthias Zwicker, Paolo Favaro, and Meiguang Jin, "Deep mean-shift priors for image restoration," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[213] Siyuan Wang, Junjie Lv, Zhuonan He, Dong Liang, Yang Chen, Minghui Zhang, and Qiegen Liu, "Denoising auto-encoding priors in undecimated wavelet domain for mr image reconstruction," Neurocomputing, vol. 437, pp. 325–338, 2021.

[214] Christian F Baumgartner, Lisa M Koch, Marc Pollefeys, and Ender Konukoglu, "An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation," in International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, 2017, pp. 111–119.

[215] N. Tustison, B. Avants, P. Cook, Y. Zheng, A. Egan, P. Yushkevich, and J. Gee, "N4itk: improved n3 bias correction," IEEE transactions on medical imaging, vol. 29, no. 6, pp. 1310–1320, 2010.

[216] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010, pp. 807–814.

[217] Lee R Dice, "Measures of the amount of ecologic association between species," Ecology, vol. 26, no. 3, pp. 297–302, 1945.

[218] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge, "Comparing images using the hausdorff distance," IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 9, pp. 850–863, 1993.

[219] Patrice Y Simard, Dave Steinkraus, and John C Platt, "Best practices for convolutional neural networks applied to visual document analysis," in Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. IEEE Computer Society, 2003, vol. 3, pp. 958–958.

[220] Thomas Varsavsky, Mauricio Orbes-Arteaga, Carole H Sudre, Mark S Graham, Parashkev Nachev, and M Jorge Cardoso, "Test-time unsupervised domain adaptation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2020, pp. 428–436.

[221] Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent, "Generalized denoising auto-encoders as generative models," Advances in Neural Information Processing Systems, vol. 26, 2013.

[222] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu, "Phiseg: Capturing uncertainty in medical image segmentation," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2019, pp. 119–127.

[223] Yan Wu, Yajun Ma, Youngwook Kee, Nataliya Kovalchuk, Dante Capaldi, Hongyi Ren, Steven Hancock, Eric Chang, Marcus Alley, John Pauly, et al., "Quantitative parametric mapping of tissues properties from standard magnetic resonance imaging enabled by deep learning," arXiv preprint arXiv:2108.04912, 2021.

[224] Neerav Karani, Georg Brunner, Ertunc Erdil, Simin Fei, Kerem Tezcan, Krishna Chaitanya, and Ender Konukoglu, "A field of experts prior for adapting neural networks at test time," arXiv preprint arXiv:2202.05271, 2022.

[225] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in International Conference on Learning Representations, 2019.

[226] S. Roth and M. J. Black, "Fields of experts: A framework for learning image priors," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005.

[227] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," IEEE Transactions on pattern analysis and machine intelligence, , no. 6, pp. 721–741, 1984.

[228] G. Hinton, "Training products of experts by minimizing contrastive divergence," Neural computation, vol. 14, no. 8, pp. 1771–1800, 2002.

[229] M. Welling, S. Osindero, and G. Hinton, "Learning sparse topographic representations with products of student-t distributions," Advances in neural information processing systems, 2002.

[230] Z. Botev and A. Ridder, "Variance reduction," Wiley StatsRef: Statistics Reference Online, pp. 1–6, 2014.

[231] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," The London, Edinburgh, and Dublin philosophical magazine and journal of science, vol. 2, no. 11, pp. 559–572, 1901.

[232] H. Hotelling, "Analysis of a complex of statistical variables into principal components," Journal of educational psychology, vol. 24, no. 6, pp. 417, 1933.

[233] D. Scott, R. Tapia, and J. Thompson, "Kernel density estimation revisited," Nonlinear Analysis: Theory, Methods & Applications, vol. 1, no. 4, pp. 339–372, 1977.

[234] R. Davis, K. S. Lii, and D. Politis, "Remarks on some nonparametric estimates of a density function," in Selected Works of Murray Rosenblatt. Springer, 2011.

[235] E. Parzen, "On estimation of a probability density function and mode," The annals of mathematical statistics, vol. 33, no. 3, pp. 1065–1076, 1962.

[236] C. Eastwood, I. Mason, C. Williams, and B. Schölkopf, "Source-free adaptation to measurement shift via bottom-up feature restoration," in 10th International Conference on Learning Representations (ICLR), Apr. 2022.

[237] JM Bland and DG Altman, "Multiple significance tests: the bonferroni method.," BMJ: British Medical Journal, vol. 310, no. 6973, pp. 170, 1995.

[238] A. Dalca, G. Balakrishnan, J. Guttag, and M. Sabuncu, "Unsupervised learning for fast probabilistic diffeomorphic registration," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018.

[239] M. Ishii and M. Sugiyama, "Source-free domain adaptation via distributional alignment by matching batch normalization statistics," arXiv preprint arXiv:2101.10842, 2021.

[240] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE international conference on computer vision, 2015.

[241] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," Advances in Neural Information Processing Systems, vol. 27, pp. 3320–3328, 2014.

[242] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "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.

[243] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2014, pp. 806–813.

[244] LI Xuhong, Yves Grandvalet, and Franck Davoine, "Explicit inductive bias for transfer learning with convolutional networks," in International Conference on Machine Learning. PMLR, 2018, pp. 2825–2834.

[245] Lilla Zöllei, Juan Eugenio Iglesias, Yangming Ou, P Ellen Grant, and Bruce Fischl, "Infant freesurfer: An automated segmentation and surface extraction pipeline for t1-weighted neuroimaging data of infants 0–2 years," Neuroimage, vol. 218, pp. 116946, 2020.

[246] George Yiasemis, Chaoping Zhang, Clara I Sánchez, Jan-Jakob Sonke, and Jonas Teuwen, "Deep mri reconstruction with radial subsampling," arXiv preprint arXiv:2108.07619, 2021.

[247] Jing Liu and David Saloner, "Accelerated mri with circular cartesian undersampling (circus): a variable density cartesian sampling strategy for compressed sensing and parallel imaging," Quantitative imaging in medicine and surgery, vol. 4, no. 1, pp. 57, 2014.

[248] Sanghyun Son, Jaeha Kim, Wei-Sheng Lai, Ming-Hsuan Yang, and Kyoung Mu Lee, "Toward real-world super-resolution via adaptive downsampling models," IEEE transactions on pattern analysis and machine intelligence, 2021.

[249] Pierre Gravel, Gilles Beaudoin, and Jacques A De Guise, "A method for modeling noise in medical images," IEEE Transactions on medical imaging, vol. 23, no. 10, pp. 1221–1232, 2004.

[250] Charles M Stein, "Estimation of the mean of a multivariate normal distribution," The annals of Statistics, pp. 1135–1151, 1981.

[251] Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu, "Mr image reconstruction using deep density priors," IEEE transactions on medical imaging, vol. 38, no. 7, pp. 1633–1642, 2018.

[252] Can Zhao, Blake E Dewey, Dzung L Pham, Peter A Calabresi, Daniel S Reich, and Jerry L Prince, "Smore: A self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning," IEEE transactions on medical imaging, vol. 40, no. 3, pp. 805–817, 2020.

[253] Cheng Peng, S Kevin Zhou, and Rama Chellappa, "Da-vsr: Domain adaptable volumetric super-resolution for medical images," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2021, pp. 75–85.

[254] Maxim Zaitsev, Julian Maclaren, and Michael Herbst, "Motion artifacts in mri: A complex problem with many partial solutions," Journal of Magnetic Resonance Imaging, vol. 42, no. 4, pp. 887–901, 2015.

[255] Benedikt Lorch, Ghislain Vaillant, Christian Baumgartner, Wenjia Bai, Daniel Rueckert, and Andreas Maier, "Automated detection of motion artefacts in mr imaging using decision forests," Journal of medical engineering, vol. 2017, 2017.

[256] Ilkay Oksuz, Bram Ruijsink, Esther Puyol-Antón, James R Clough, Gastao Cruz, Aurelien Bustin, Claudia Prieto, Rene Botnar, Daniel Rueckert, Julia A Schnabel, et al., "Automatic cnn-based detection of cardiac mr motion artefacts using k-space data augmentation and curriculum learning," Medical image analysis, vol. 55, pp. 136–147, 2019.

[257] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[258] Diederik P Kingma and Max Welling, "An introduction to variational autoencoders," arXiv preprint arXiv:1906.02691, 2019.

[259] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis, "Compressed sensing using generative models," in International Conference on Machine Learning. PMLR, 2017, pp. 537–546.

[260] Razvan V Marinescu, Daniel Moyer, and Polina Golland, "Bayesian image reconstruction using deep generative models," arXiv preprint arXiv:2012.04567, 2020.

[261] Niklas Smedemark-Margulies, Jung Yeon Park, Max Daniels, Rose Yu, Jan-Willem van de Meent, and Paul Hand, "Generator surgery for compressed sensing," arXiv preprint arXiv:2102.11163, 2021.

[262] Reinhard Heckel and Paul Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," in International Conference on Learning Representations, 2019.

[263] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu, "Transferring gans: generating images from limited data," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 218–234.

[264] Masataka Yamaguchi, Yuma Koizumi, and Noboru Harada, "Adaflow: Domain-adaptive density estimator with application to anomaly detection and unpaired cross-domain translation," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3647–3651.

[265] Yonina C Eldar, "Generalized sure for exponential families: Applications to regularization," IEEE Transactions on Signal Processing, vol. 57, no. 2, pp. 471–481, 2008.

[266] Magauiya Zhussip, Shakarim Soltanayev, and Se Young Chun, "Extending stein's unbiased risk estimator to train deep denoisers with correlated pairs of noisy images," Advances in neural information processing systems, vol. 32, pp. 1465–1475, 2019.

[267] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila, "Noise2noise: Learning image restoration without clean data," arXiv preprint arXiv:1803.04189, 2018.

[268] Hemant Kumar Aggarwal and Mathews Jacob, "Model adaptation for image reconstruction using generalized stein's unbiased risk estimator," arXiv preprint arXiv:2102.00047, 2021.

[269] Robert A Vandermeulen, René Saitenmacher, and Alexander Ritchie, "A proposal for supervised density estimation," 2020.

[270] Michael Gutmann and Aapo Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[271] Pascal Vincent, "A connection between score matching and denoising autoencoders," Neural computation, vol. 23, no. 7, pp. 1661–1674, 2011.

[272] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll, "Learning a variational network for reconstruction of accelerated mri data," Magnetic resonance in medicine, vol. 79, no. 6, pp. 3055–3071, 2018.

List of Publications

Articles in peer-reviewed journals

- Karani, Erdil, Chaitanya, Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis, 2021.
- Tezcan, Karani, Baumgartner, Konukoglu. "Sampling possible reconstructions of undersampled acquisitions in MR imaging with a deep learned prior." IEEE Transactions on Medical Imaging, 2022.
- Chaitanya, Karani, Baumgartner, Erdil, Becker, Donati, Konukoglu. Semi-supervised task-driven data augmentation for medical image segmentation. Medical Image Analysis, 2021.
- Karani, Zhang, Tanner, Konukoglu. An image interpolation approach for acquisition time reduction in navigator-based 4D MRI. Medical image analysis, 2019.
- Karani, Tanner, Kozerke, Konukoglu. Reducing navigators in free-breathing abdominal MRI via temporal interpolation using convolutional neural networks. IEEE transactions on medical imaging, 2018.
- Ozdemir, Karani, Fürnstahl, Goksel. Interactive segmentation in MRI for orthopedic surgery planning: bone tissue. International Journal of Computer Assisted Radiology and Surgery, 2017.

Conference contributions

- Erdil, Chaitanya, Karani, Konukoglu. Task-agnostic out-of-distribution detection using kernel density estimation. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging Workshop, MICCAI, 2021.
- Chaitanya, Erdil, Karani, Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. In Neural Information Processing System, 2020.
- Volokitin, Erdil, Karani, Tezcan, Chen, Van Gool, Konukoglu. Modelling the distribution of 3D brain MRI using a 2D slice VAE. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020.
- Chaitanya, Karani, Baumgartner, Becker, Donati, Konukoglu. Semi-supervised and task-driven data augmentation. In International conference on information processing in medical imaging, 2019.
- Karani, Chaitanya, Baumgartner, Konukoglu. A lifelong learning approach to brain MR segmentation across scanners and protocols. In International Con-

ference on Medical Image Computing and Computer-Assisted Intervention, 2018.

- Zhang, Karani, Tanner, Konukoglu. Temporal interpolation via motion field prediction. In Medical Imaging with Deep Learning, 2018.
- Karani, Tanner, Kozerke, Konukoglu. Temporal interpolation of abdominal MRIs acquired during free-breathing. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017.

Preprints

- Karani, Brunner, Fei, Tezcan, Erdil, Chaitanya, Konukoglu. A field-of-experts prior for adapting neural networks at test time, arXiv preprint arXiv:2202.05271, 2022.

Curriculum Vitae

Personal data

Name	Neerav Karani
Date of Birth	October 31, 1989
Place of Birth	Patna, Bihar, India
Citizen of	India

Education

2017 – 2022	Doctor of Science	ETH Zurich, Switzerland
2015 – 2017	Master of Science	ETH Zurich, Switzerland
2012 – 2013	Master of Technology	Indian Institute of Technology Madras, India
2008 – 2012	Bachelor of Technology	Indian Institute of Technology Madras, India

Work Experience

2017 – 2022	Scientific Assistant	ETH Zurich, Switzerland
2013 – 2015	Senior Electrical Engineer	Philips Healthcare, India