# Advancing brain MRI assessment using generative models

by

Farzad BEIZAEE

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, SEPTEMBER 10, 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

**BOARD OF EXAMINERS**

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. José Dolz, Thesis supervisor
Department of Software Engineering and IT, École de technologie supérieure

Mr. Christian Desrosiers, Thesis Co-Supervisor
Department of Software Engineering and IT, École de technologie supérieure

Mr. Gregory A. Lodygensky, Thesis Co-Supervisor
Research Center, CHU Sainte-Justine

Mr. Jean-Marc Lina, Chair, Board of Examiners
Department of Electrical Engineering, École de technologie supérieure

Mr. Carlos Vàzquez, Member of the Jury
Department of Software Engineering and IT, École de technologie supérieure

Mr. Mahdi Hosseini, External Independent Examiner
Department of Computer Science and Software Engineering, Concordia University

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON AUGUST 27, 2025

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

## ACKNOWLEDGEMENTS

This thesis marks the end of a significant chapter in my life, and it would not have been possible without the support, kindness, and belief of many incredible people to whom I am deeply grateful.

First, I wish to express my sincere gratitude to my main supervisor, Prof. José Dolz, whose exceptional knowledge, insightful guidance, and continuous support have been invaluable through all stages of this journey. From the very beginning, Jose has been more than just a supervisor; he has been a mentor, a supporter, and a friend. I also sincerely thank my co-supervisor, Prof. Christian Desrosiers, not only for his support and guidance but also for the genuinely kind and thoughtful person he is. His patience, attentiveness, and constructive feedback made every interaction both insightful and encouraging. I am also thankful to Dr. Gregory Lodygensky, a great doctor and my clinical co-supervisor, for his clinical insights, boundless energy, and the belief he had in me. Gregory's dedication to his work, enthusiasm and passion for research were always inspiring. I could not be more grateful to have worked under the supervision of such brilliant people. I would also like to thank my friends and lab-mates for the stimulating discussions, collaboration, and friendly atmosphere that made this journey more enjoyable.

My deepest appreciation belongs to my family. My heartfelt thanks go to my beloved wife, Hanieh, who was there for me through every up and down, late night, and stressful deadline. Her love, patience, and unwavering support made not only this PhD but also life itself a lighter and more beautiful journey. I also want to thank my father and mother, whose unconditional love, sacrifices, and endless belief in me laid the very foundation upon which everything else stands. Words will never suffice to express my gratitude for all they have given me.

Finally, I extend my appreciation to all those whose paths crossed mine during these years, leaving behind invaluable lessons and cherished memories.

**Advancing brain MRI assessment using generative models**

Farzad BEIZAEE

## ABSTRACT

Magnetic Resonance Imaging (MRI) has become an indispensable modality in neurological diagnostics, offering non-invasive, high-resolution insights into brain structure and function. Despite its clinical ubiquity, automated analysis of brain MRI faces persistent challenges, including domain variability across imaging sites, limited annotated data for target tasks, and the subtlety of certain developmental and pathological anomalies. Addressing these limitations is crucial for advancing diagnostic reliability, generalization, and scalability in real-world clinical settings.

This thesis explores the integration of generative models into brain MRI analysis as a principled and label-efficient alternative to traditional discriminative frameworks. By modeling the underlying distribution of brain MRI data, generative approaches provide robust tools for learning from unlabeled data, simulating anatomical variability, and enhancing interpretability. We focus on three key areas of application: unsupervised MRI harmonization, unsupervised brain anomaly detection, and neonatal brain age estimation.

First, we introduce Harmonizing Flows, a novel framework based on normalizing flows for unsupervised and source-free harmonization of multi-site MRI scans. This method effectively aligns data distributions across scanners while preserving clinically relevant features, significantly improving model generalization, even on unseen domains. Second, we leverage a progression of generative models and propose three unsupervised anomaly detection approaches, MAD-AD, DeCo-Diff, and REFLECT, each building upon the strengths and addressing the shortcomings of the previous, leading to progressively more robust and effective solutions. In particular, these models learn the manifold of healthy brain anatomy and isolate pathological deviations without requiring annotated anomalies, demonstrating strong performance across various datasets. Lastly, we propose a learning-based framework for predicting neonatal brain age, enabling the identification of infants at risk of neurodevelopmental delays by quantifying maturational discrepancies not evident in conventional structural assessments.

Together, these contributions establish a cohesive, generative-model-based framework for brain MRI assessment that is scalable, interpretable, and clinically meaningful. Through extensive evaluation across diverse populations and imaging conditions, the proposed methods demonstrate improved robustness, enhanced diagnostic capability, and cross-domain generalizability. This work underscores the transformative potential of generative models in neuroimaging and paves the way toward more accessible, equitable, and effective brain health diagnostics.

**Keywords:** Brain MRI assessment, generative models, unsupervised MRI harmonization, unsupervised anomaly detection, neonatal brain age estimation

**Évaluation par IRM cérébrale à l'aide de modèles génératifs**

Farzad BEIZAEE

## RÉSUMÉ

L'imagerie par résonance magnétique (IRM) est une modalité essentielle du diagnostic neurologique, offrant une visualisation non invasive et haute résolution de la structure et du fonctionnement cérébral. Malgré son adoption généralisée, l'analyse automatisée de l'IRM se heurte à des défis persistants : variabilité entre sites, rareté des annotations, et subtilité de certaines anomalies développementales ou pathologiques. Surmonter ces limites est crucial pour renforcer la fiabilité, la généralisation et la mise à l'échelle du diagnostic en contexte clinique réel.

Cette thèse explore l'intégration des modèles génératifs dans l'analyse des IRM cérébrales comme alternative rigoureuse et économe en annotations aux approches discriminantes classiques. En modélisant la distribution sous-jacente des données d'IRM cérébrales, ces modèles offrent des outils robustes pour l'apprentissage non supervisé, la simulation de la variabilité anatomique et une meilleure interprétabilité. Trois domaines d'application principaux sont étudiés : l'harmonisation non supervisée des IRM cérébrales, la détection non supervisée d'anomalies cérébrales, et l'estimation de l'âge cérébral néonatal.

Nous introduisons d'abord Harmonizing Flows, un nouveau cadre basé sur les normalizing flows pour l'harmonisation non supervisée et sans accès aux données sources des IRM multi-sites. Cette méthode aligne efficacement les distributions entre scanners tout en préservant les caractéristiques cliniques, améliorant ainsi la généralisation, même sur des domaines non vus. Ensuite, nous présentons une progression de modèles génératifs pour la détection d'anomalies non supervisée : MAD-AD, DeCo-Diff et REFLECT. Chacune s'appuie sur les forces de la précédente et corrige ses limites, menant à des solutions plus robustes. Ces modèles apprennent le *variété* de l'anatomie cérébrale saine et identifient les déviations pathologiques sans annotations, avec d'excellents résultats sur divers jeux de données. Enfin, nous proposons une méthode pour estimer l'âge cérébral néonatal, permettant d'identifier les nourrissons à risque de retards en quantifiant des écarts de maturation invisibles aux évaluations classiques.

Ensemble, ces contributions établissent un cadre cohérent d'analyse IRM cérébrale fondé sur des modèles génératifs, alliant évolutivité, interprétabilité et utilité clinique. Des évaluations sur des populations et conditions variées démontrent une robustesse, une capacité diagnostique accrue et une généralisation inter-domaines renforcée. Ce travail met en lumière le potentiel transformateur des modèles génératifs en neuro-imagerie et ouvre la voie à des diagnostics plus accessibles, équitables et efficaces.

**Mots-clés:** Évaluation de l'IRM cérébrale, modèles génératifs, harmonisation non supervisée des IRM, détection non supervisée d'anomalies, estimation de l'âge cérébral néonatal

## TABLE OF CONTENTS

Page


INTRODUCTION ..... 1
0.1 Brain Magnetic Resonance Imaging ..... 1
0.1.1 Magnetic Resonance Imaging Principles and Properties ..... 2
0.1.2 Clinical Applications of Brain MRI ..... 4
0.2 Brain MRI Assessment ..... 6
0.2.1 Traditional approaches for brain MRI assessment. ..... 6
0.2.2 Recent approaches using Deep Neural Networks. ..... 7
0.3 Challenges and Motivation ..... 8
0.4 Generative Models for Brain MRI Assessment: Rationale and Potential ..... 11
0.5 Proposed Solutions and Research Objectives ..... 13
0.5.1 MRI Harmonization ..... 13
0.5.2 Unsupervised Anomaly Detection ..... 14
0.5.3 Neonatal Brain Age Estimation ..... 15
0.6 Contributions ..... 16
0.7 Integration and Broader Impact ..... 19

CHAPTER 1 LITERATURE REVIEW ..... 21
1.1 Preliminaries: Generative models ..... 21
1.1.1 Normalizing Flows ..... 21
1.1.2 Diffusion Models ..... 23
1.1.3 Rectified Flows ..... 25
1.2 MRI harmonization ..... 27
1.2.1 Traditional Statistical Methods ..... 27
1.2.2 Learning-Based Harmonization Approaches ..... 28
1.3 Unsupervised Anomaly Detection ..... 33
1.3.1 Embedding-Based UAD Methods ..... 34
1.3.2 Synthetic-Based UAD Methods ..... 35
1.3.3 Reconstruction-Based UAD Methods ..... 36
1.3.3.1 Reconstruction-Based UAD Methods Using VAEs ..... 37
1.3.3.2 Reconstruction-Based UAD Methods using GANs ..... 39
1.3.3.3 Reconstruction-Based UAD Methods using Diffusion Models . 40
1.4 Neonatal Brain Age Estimation ..... 43

CHAPTER 2 HARMONIZING FLOWS: LEVERAGING NORMALIZING FLOWS FOR UNSUPERVISED AND SOURCE-FREE MRI HARMONIZATION ..... 47
2.1 Introduction ..... 47
2.2 Related work ..... 49
2.3 Methodology ..... 54
2.3.1 Learning the source domain distribution ..... 55

2.3.2 Achieving image harmonization ..... 58
2.4 Experiments ..... 60
2.4.1 Experimental setting ..... 60
2.4.2 Results ..... 65
2.4.2.1 Performance on the segmentation task ..... 66
2.4.2.2 Performance on neonatal brain age estimation ..... 76
2.4.2.3 A closer look at the harmonization performance ..... 78
2.4.2.4 Structural integrity. ..... 81
2.4.2.5 Friedman Ranking ..... 82
2.5 Conclusion ..... 84

CHAPTER 3 MAD-AD: MASKED DIFFUSION FOR UNSUPERVISED BRAIN ANOMALY DETECTION ..... 87
3.1 Introduction ..... 87
3.2 Related works ..... 89
3.3 Method ..... 92
3.3.1 Modeling the normal feature space ..... 92
3.3.2 Recovering normal images ..... 95
3.3.3 Anomaly localization ..... 96
3.4 Experiments ..... 97
3.4.1 Experimental setting ..... 97
3.4.2 Results ..... 99
3.5 Conclusion ..... 101

CHAPTER 4 CORRECTING DEVIATIONS FROM NORMALITY: A REFORMULATED DIFFUSION MODEL FOR MULTI-CLASS UNSUPERVISED ANOMALY DETECTION ..... 103
4.1 Introduction ..... 103
4.2 Related work ..... 106
4.2.1 Unsupervised Anomaly detection ..... 106
4.2.2 Multi-class unsupervised anomaly detection ..... 107
4.3 Preliminaries ..... 108
4.3.1 Forward Diffusion Process ..... 109
4.3.2 Reverse Diffusion Process ..... 109
4.4 Method ..... 111
4.4.1 Modeling anomalies as noise in latent space ..... 111
4.4.2 Deviation instead of Diffusion ..... 112
4.4.3 Anomaly detection ..... 115
4.5 Experiments ..... 117
4.5.1 Setting ..... 117
4.5.2 Results ..... 119
4.5.2.1 Quantitative Results ..... 119
4.5.2.2 Qualitative Results ..... 119
4.5.2.3 Ablation Studies ..... 120

4.6 Conclusions ..... 121

CHAPTER 5 REFLECT: RECTIFIED FLOWS FOR EFFICIENT BRAIN ANOMALY CORRECTION TRANSPORT ..... 123
5.1 Introduction ..... 123
5.2 Preliminaries: Rectified Flows ..... 126
5.3 Method ..... 127
5.3.1 Generating Paired Samples ..... 128
5.3.2 Training Rectified Flows for Anomaly Correction ..... 129
5.3.3 Inference and Anomaly Localization ..... 130
5.4 Experiments ..... 130
5.4.1 Experimental Setting ..... 130
5.4.2 Results ..... 131
5.5 Conclusion ..... 133

CHAPTER 6 DETERMINING REGIONAL BRAIN GROWTH IN PREMATURE AND MATURE INFANTS IN RELATION TO AGE AT MRI USING DEEP NEURAL NETWORKS ..... 135
6.1 Introduction ..... 135
6.2 Results ..... 138
6.2.1 Main results ..... 138
6.2.2 On the importance of different regions ..... 140
6.2.3 Comparison to existing brain age bio-markers ..... 141
6.2.4 Performance on low labeled data regime ..... 142
6.2.5 The impact of different backbones ..... 143
6.3 Discussion ..... 145
6.3.1 Limitations ..... 148
6.4 Materials and Methods ..... 148
6.4.1 Dataset ..... 149
6.4.2 Methodology ..... 149
6.4.2.1 Brain segmentation ..... 150
6.4.2.2 Extracting representative features ..... 150
6.4.2.3 Postmenstrual age regression ..... 152
6.4.2.4 Discovering important correlations ..... 153
6.4.3 Evaluation protocol ..... 153
6.4.4 Compared methods ..... 154
6.4.5 Implementation details ..... 154

CONCLUSION AND RECOMMENDATIONS ..... 155

BIBLIOGRAPHY ..... 159

## LIST OF TABLES

Page

Table 2.1 **Acquisition parameters across different sites.** Scanner details and phenotypic information for each site used in this study. *y*: years; *gw*: gestation weeks ..... 60

Table 2.2 **Performance overview on the cross-site adult MRI segmentation task.** Segmentation performance, in terms of DSC and HD95 metrics, across different harmonization approaches. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold** ..... 63

Table 2.3 **Performance overview on the cross-site neonatal MRI segmentation task.** Segmentation performance, in terms of DSC and HD95 metrics, across different harmonization approaches. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold** ..... 68

Table 2.4 **Performance overview on the cross-site neonatal age estimation task.** Brain age estimation performance, in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics, across different harmonization approaches and modalities. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold** ..... 77

Table 3.1 **Performance in setting *S1:*** results across different lesion sizes, where bold highlights the best method and improvements of our approach compared to the best baseline are indicated in green ..... 99

Table 3.2 **Performance in setting *S2:*** results across different modalities, where bold highlights the best method and performance improvements (*resp.* decrease) of our approach compared to the best baseline are indicated in **green** (*resp.* **red**) ..... 100

Table 3.3 Effect of different sources for the anomaly score in MAD-AD (BRATS'21) ..... 101

Table 3.4 Ablation study on two key hyper-parameters of MAD-AD ..... 101

Table 4.1 Quantitatve evaluation of DeCo-Diff on MVTec-AD. ..... 114

Table 4.2 Quantitatve evaluation of DeCo-Diff on VisA. ..... 117

Table 4.3 Impact of the correction strategy at each time-step and number of reverse steps ..... 121

Table 4.4 Different levels of discrepancy. Results on MVTec-AD ..... 122

Table 5.1 Quantitative results of REFLECT. ..... 131

Table 5.2 Ablation studies on model size and VAE model employed for REFLECT. 133

Table 6.1 Quantitative results compared to state-of-the-art learning-based brain age estimation methods ..... 139

Table 6.2 Comparison of the quantitative performance obtained by the proposed features and existing features for brain age estimation ..... 142

Table 6.3 Quantitative performance of different brain age estimation methods based on a fraction of labeled data ..... 143

Table 6.4 Results using different segmentation and regression backbones ..... 145

## LIST OF FIGURES


Page

Figure 0.1 Schematic overview of brain magnetic resonance imaging ..... 2

Figure 0.2 Examples of common brain MRI sequences ..... 3

Figure 0.3 Examples of T1-weighted MRI scans from eight different sites, illustrating the variability in image contrast across acquisition settings ... 9

Figure 0.4 MRI variability across different scanners ..... 14

Figure 1.1 Normalizing flow distribution transformation ..... 22

Figure 1.2 Score-based diffusion model forward and Reverse SDE transformations . 23

Figure 2.1 Overview of Harmonizing Flows pipeline ..... 50

Figure 2.2 Cross-site brain MRI segmentation matrix across the compared methods ..... 66

Figure 2.3 Effect of each component of Harmonizing Flows ..... 70

Figure 2.4 The effect of different stopping criteria to stop the harmonizer network adaptation. ..... 71

Figure 2.5 Best stopping criteria metric. ..... 72

Figure 2.6 Ablation study on architecture components and hyper-parameters of the harmonizing flow. ..... 73

Figure 2.7 Harmonized samples across sites using the proposed method. ..... 74

Figure 2.8 Visual examples of neonatal MRI brain images harmonized using the proposed method. ..... 76

Figure 2.9 Example of harmonized images using different methods. ..... 77

Figure 2.10 Histograms of the harmonized MRIs from multiple target datasets compared to the histogram of the source MRIs ..... 79

Figure 2.11 Segmentation and Age estimation results vs. WD of intensity histograms for compared Harmonization methods. ..... 81

Figure 2.12 Visualization of harmonization applied to a sample of neonatal injured brain MRI. ..... 82

Figure 2.13 Friedman Rank for the compared harmonization methods. ..... 83

Figure 3.1 Overview of the proposed MAD-AD method. ..... 89

Figure 3.2 Visual example of the reverse process. ..... 97

Figure 3.3 MAD-AD Qualitative results. .....102

Figure 4.1 Overview of the proposed DeCo-Diff method. .....105

Figure 4.2 Diffusion model reconstruction vs. DeCo-Diff. .....106

Figure 4.3 DeCo-Diff qualitative results. .....110

Figure 4.4 Visualization of deviation correction through the reverse process steps. 120

Figure 5.1 Overview of REFLECT method. .....127

Figure 5.2 REFLECT qualitative results. .....132

Figure 5.3 Transition between an anomalous brain to its healthy counterpart using REFLECT. .....133

Figure 6.1 Overview of the proposed brain age estimation method. .....138

Figure 6.2 a) Predicted age *versus* real age; b) Prediction error for different age intervals. .....140

Figure 6.3 Histogram showing the brain regions with the best scaled PFI .....141

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| AD | Anomaly Detection |
| AE | Autoencoder |
| AI | Artificial Intelligence |
| ALS | Amyotrophic Lateral Sclerosis |
| BCE | Binary Cross-Entropy |
| BPD | Bits Per Dimension |
| BraTS | Brain Tumor Segmentation |
| CE | Cross-Entropy |
| CNF | Continuous Normalizing Flows |
| CNN | Convolutional Neural Network |
| CNR | Contrast-to-Noise Ratio |
| CSF | Cerebrospinal Fluid |
| CT | Computed Tomography |
| DAE | Denoising Autoencoder |
| DDIM | Denoising Diffusion Implicit Models |
| DDPM | Denoising Diffusion Probabilistic Models DNN: Deep Neural Network |
| DoD | Direction of Deviation |
| DTI | Diffusion Tensor Imaging |
| ELU | Exponential Linear Unit |

| | |
|---|---|
| ETS | École de Technologie Supérieure |
| FA | Fractional Anisotropy |
| FLAIR | Fluid-Attenuated Inversion Recovery |
| GAN | Generative Adversarial Network |
| GM | Gray Matter |
| IUGR | Intrauterine Growth Restriction |
| KL | Kullback-Leibler |
| KLD | Kullback-Leibler Divergence |
| MS | Multiple Sclerosis |
| MRI | Magnetic Resonance Imaging |
| NF | Normalizing Flows |
| PD | Proton Density |
| PET | Positron Emission Tomography |
| PFI | Permutation Feature Importance |
| PMA | Postmenstrual Age |
| ODE | Ordinary Differential Equation |
| RELU | Rectified Linear Unit |
| RV | Relational Volume |
| SDE | stochastic differential equation |
| SFDA | Source-Free Domain Adaptation |

| | |
|---|---|
| SSIM | Structural Similarity Index Measure |
| SVR | Surface-to-Volume Ratio |
| T1-w | T1-weighted |
| T1CE | T1-Weighted Contrast-Enhanced |
| T2-w | T2-weighted |
| TBI | Traumatic Brain Injuries |
| TTA | Test-Time Adaptation |
| UAD | Unsupervised Anomaly Detection |
| UDA | Unsupervised Domain Adaptation |
| VAE | Variational AutoEncoder |
| WD | Wasserstein Distance |
| WM | White Matter |

## LIST OF SYMBOLS AND UNITS OF MEASUREMENTS

| | |
|---|---|
| AP | Average Precision |
| AUPRC | Area Under the Precision-Recall Curve |
| AUROC | Area Under the Receiver Operating Characteristic Curve |
| AUPRO | Area Under the PRO curve |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |
| $R^2$ | determination Score |
| DSC | Dice Similarity Coefficient |
| HD | Hausdroff Distance |
| JD | Jaccard Distance |
| $\alpha$ | Learning-rate or general weighting coefficient |
| $\beta$ | Momentum, scaling, or precision parameter |
| $\gamma$ | Scale or discount factor |
| $\lambda$ | Regularisation / penalty weight |
| $\eta$ | Step-size or efficiency parameter |
| $\theta$ | Model or function parameters |
| $\phi$ | Feature-mapping or activation-function parameters |
| $\chi$ | variable |

| | |
|---|---|
| $\tau$ | Time constant or temperature parameter |
| $\boldsymbol{x}_i$ | $i^{\text{th}}$ input sample |
| $\boldsymbol{z}_i$ | $i^{\text{th}}$ latent sample |
| $\Omega$ | Spatial domain (e.g. image width $\times$ height) |
| $p_x(x)$ | Probability distribution of input data $x$ |
| $u \sim p_u(u)$ | Latent variable drawn from a base distribution $p_u$ (often $\mathcal{N}(0, 1)$) |
| $\mathcal{L}$ | Loss function |
| $\mathbb{E}$ | Expectation operator |
| $D$ | Domain |
| $\mathcal{X}$ | Input space |
| $\mathcal{Y}$ | Label / target space |

# INTRODUCTION

## 0.1 Brain Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) has become one of the most vital tools in modern medical diagnostics, revolutionizing the way clinicians and researchers visualize, understand, and diagnose neurological conditions (Brown *et al.*, 2014). This non-invasive imaging technique leverages powerful magnetic fields, radiofrequency pulses, and computational reconstruction algorithms to generate high-resolution images of brain structures, offering unparalleled insights into both normal anatomy and pathological alterations (Lin, 2000; Dale, Brown & Semelka, 2015). The exceptional contrast resolution of MRI provides a comprehensive assessment of brain health by enabling the differentiation between soft tissues, grey and white matter, cerebrospinal fluid, and various pathological entities such as tumors, lesions, and vascular abnormalities, which are typically difficult or impossible to detect with other imaging modalities (van der Graaf, 2010).

Moreover, MRI's capacity to reveal subtle changes in tissue structure or function, often before clinical symptoms emerge, makes it particularly valuable for early diagnosis. The value of early detection through MRI lies not only in diagnostic clarity but also plays a pivotal role in guiding timely interventions that can significantly improve patient outcomes (Frisoni *et al.*, 2010; Aldossary, Kotb & Kamal, 2019). For many neurological conditions, early MRI findings enable critical decisions that help halt disease progression or reduce long-term damage (Landfeldt *et al.*, 2018; Schwarz, 2021). Also, clinical studies have shown that prompt actions based on MRI can substantially reduce the risk of death or severe disability (Lee *et al.*, 2021; Kumar *et al.*, 2024).

Furthermore, in pediatric populations, particularly among preterm infants, MRI plays a crucial role in assessing brain development. Notably, the rate of preterm births has been steadily increasing worldwide, posing an escalating public health concern (Blencowe *et al.*, 2013).

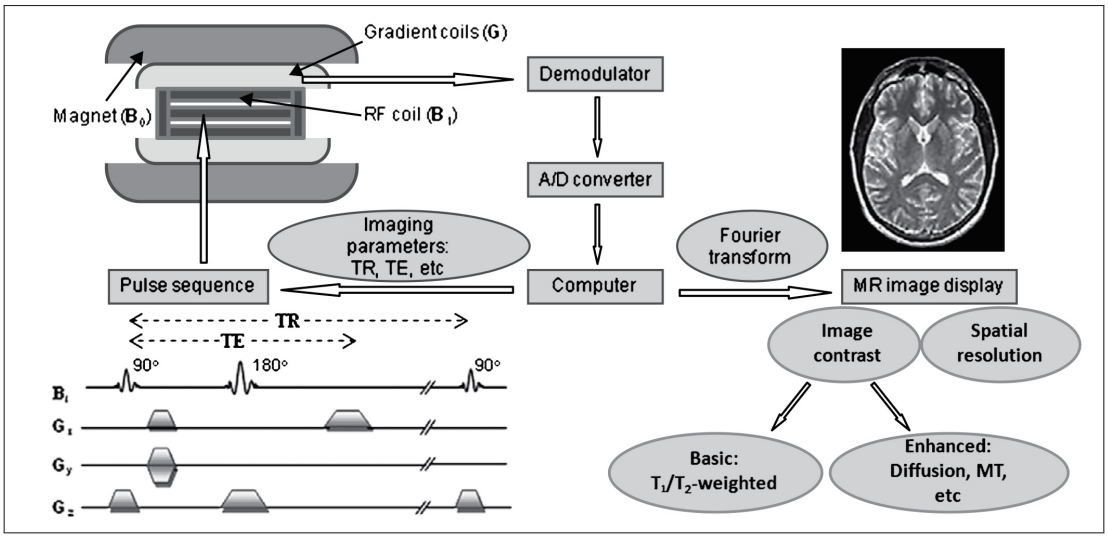



Figure 0.1 Schematic overview of brain magnetic resonance
Taken from Xu *et al.* (2008)

Preterm birth complications are the leading cause of death among children under five years of age, and most of them are because of inappropriate brain growth or brain injuries. However, three-quarters of these deaths could be prevented with current interventions (Liu *et al.*, 2016). Therefore, brain MRI not only enhances diagnostic precision but also serves as a critical enabler of life-saving interventions, underscoring its pivotal role in modern neuroimaging and public health.

### 0.1.1 Magnetic Resonance Imaging Principles and Properties

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that uses strong magnetic fields and radiofrequency waves to visualize internal structures of the body, particularly soft tissues like the brain. At the core of MRI lies the behavior of hydrogen atoms, which are abundant in water and fat throughout the human body. When a person is placed inside a powerful MRI scanner, the hydrogen nuclei align with the magnetic field. A brief radiofrequency pulse then disturbs this alignment. As the nuclei return to their original state, they emit signals that are captured by the scanner and processed into detailed images (Lin, 2000; Brown *et al.*, 2014;

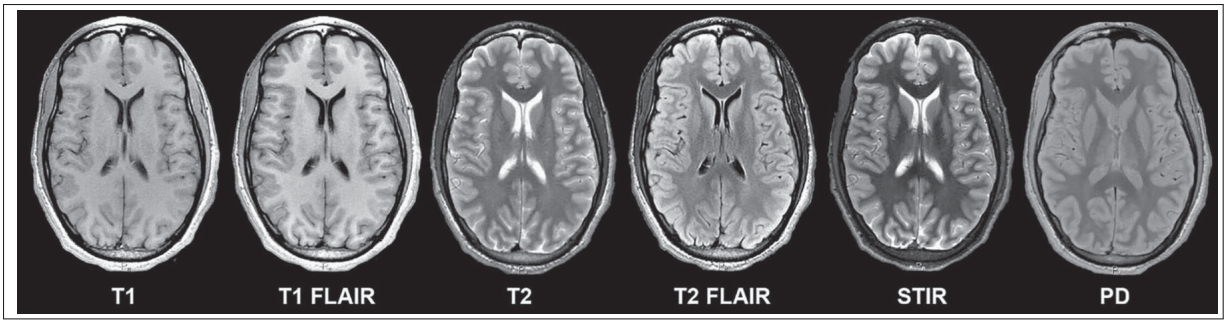



Figure 0.2 Examples of common brain MRI sequences, including T1-weighted, T1-FLAIR, T2-weighted, T2-FLAIR, STIR, and proton density (PD)
Taken from Tanenbaum *et al.* (2017)

Dale *et al.*, 2015). A schematic overview of brain magnetic resonance imaging is visualized in Figure 0.1.

What makes MRI uniquely suited for brain imaging is its ability to differentiate between tissues based on how quickly these hydrogen atoms return to equilibrium, known as relaxation. Different tissues, such as gray matter, white matter, and cerebrospinal fluid, relax at different rates, producing distinct signal intensities that are used to generate contrast in the images (Lin, 2000; Brown *et al.*, 2014). By modifying the timing and characteristics of the radiofrequency pulses, MRI can emphasize different tissue properties. This results in a range of imaging sequences, such as T1-weighted, T2-weighted, and diffusion-weighted imaging, each tailored to highlight specific features or pathologies (Bitar *et al.*, 2006). Different brain MRI sequences are showcased in Figure 0.2.

MRI also offers several technical advantages. First, it provides high spatial resolution and excellent soft-tissue contrast without using ionizing radiation, making it particularly safe for repeated use in both adults and children. Moreover, MRI can go beyond static anatomy. Indeed, advanced techniques like diffusion imaging (Le Bihan, 2003) and functional MRI (Heeger & Ress, 2002) can capture microstructural integrity or brain activity, providing deeper insights into both health and disease.

### 0.1.2 Clinical Applications of Brain MRI

One primary advantage of MRI in clinical practice is its ability to detect brain injuries both effectively and accurately. Brain injuries, whether traumatic or non-traumatic, exhibit diverse imaging characteristics. MRI provides a highly sensitive assessment for a wide spectrum of traumatic brain injuries (TBI), ranging from mild concussions to severe diffuse axonal injuries. Unlike computed tomography (CT), MRI can detect subtle abnormalities such as microhemorrhages, contusions, and diffuse axonal injuries, even in cases where CT scans appear normal (Smith, Hicks & Povlishock, 2013; Shenton *et al.*, 2012). This sensitivity makes MRI indispensable for evaluating both acute and chronic stages of brain injury, guiding clinical management, and informing prognoses.

Beyond traumatic injuries, MRI is also essential for diagnosing and monitoring numerous non-traumatic neurological disorders. For progressive neurodegenerative conditions, such as Alzheimer's disease (Jack *et al.*, 2010) and Parkinson's disease (Lehéricy *et al.*, 2012), MRI provides detailed insights into structural and functional changes, aiding in early detection and disease tracking. Similarly, in conditions like multiple sclerosis (MS) (Polman *et al.*, 2011) and amyotrophic lateral sclerosis (ALS) (Turner *et al.*, 2012), MRI enables precise visualization of lesions and neural degeneration, guiding clinical management. Beyond neurodegeneration, MRI plays a critical role in identifying and characterizing vascular pathologies, such as ischemic and hemorrhagic strokes (Chalela *et al.*, 2007; Wardlaw *et al.*, 2013). Also, by offering non-invasive, high-resolution imaging of cerebral vasculature, MRI significantly reduces patient risk compared to invasive angiography techniques (Brinjikji *et al.*, 2014).

Furthermore, MRI is pivotal in neuro-oncology, particularly in the assessment of brain tumors. Its superior contrast resolution and multiparametric capabilities, such as T1-weighted, T2-weighted, FLAIR, and diffusion-weighted imaging, enable the differentiation of tumor tissue from normal brain structures, support tumor classification, and assist in evaluating malignancy and planning

surgical interventions (Pope *et al.*, 2005). For instance, gliomas, the most common type of brain tumor, can be precisely visualized, assessed, and monitored with MRI, enabling treatment planning and evaluation of therapeutic response or recurrence (Ellingson *et al.*, 2014; Wen *et al.*, 2010). Overall, MRI is vital for identifying lesions, monitoring disease progression, and evaluating treatment efficacy. These capabilities not only facilitate early diagnosis but also support timely and targeted interventions, ultimately improving patient outcomes and quality of life.

Lastly, brain MRI plays a pivotal role in neonatal neuroimaging, essential for assessing brain development, maturation, and detecting subtle anomalies such as intrauterine growth restriction (IUGR). It provides superior soft tissue contrast and allows non-invasive evaluation of both structural and functional development (Hintz, 2015). Early MRI assessment of neonates helps clinicians understand developmental trajectories, enabling timely and targeted interventions that can significantly improve long-term neurological outcomes (Dubois *et al.*, 2014). Moreover, MRI-based brain age estimation techniques provide valuable metrics by comparing predicted brain age to chronological age, helping identify infants at risk of neurodevelopmental impairment (Gholipour, 2017; Cole, 2017).

In summary, brain MRI, grounded in its sophisticated principles of magnetic resonance and versatile clinical applications, has become a cornerstone of neurological diagnostics, playing a pivotal role across diverse clinical scenarios, including injury assessment, neurodegenerative disease evaluation, vascular disorder detection, oncological imaging, and pediatric developmental assessment. Its versatility, accuracy, and non-invasive nature make MRI indispensable for identifying lesions, monitoring disease progression, and evaluating treatment efficacy. These capabilities not only facilitate early and precise diagnosis but also support timely, targeted interventions, ultimately advancing treatment planning and improving patient outcomes and quality of life.

## 0.2 Brain MRI Assessment

### 0.2.1 Traditional approaches for brain MRI assessment.

Classical brain MRI analysis has long relied on established image processing techniques and statistical modeling frameworks to extract meaningful information from structural and functional scans. These traditional methods, rooted in anatomical atlases, voxel-wise statistics, and handcrafted features, formed the backbone of neuroimaging research and clinical assessment. Widely adopted software packages such as FSL (Smith *et al.*, 2004) and FreeSurfer (Fischl, 2012) exemplify this era of analysis, offering robust, validated tools for tissue segmentation, cortical surface reconstruction, and volumetric analysis. These frameworks often rely on multi-stage pipelines involving preprocessing (e.g., skull stripping, motion correction, spatial normalization), feature extraction, and hypothesis-driven statistical modeling.

In clinical and research contexts, abnormalities such as tumors, lesions, or traumatic injuries were traditionally identified by comparing quantitative biomarkers, such as tissue volumes, cortical thickness, or diffusion metrics, against normative data from healthy populations (Ashburner & Friston, 2000). Deviations from these norms were often indicative of potential pathology. Feature-based classification was also widely used for disease prediction, where hand-engineered descriptors (e.g., texture, shape, or intensity statistics, for example) were input into machine learning models such as support vector machines, random forests, or logistic regression classifiers (Klöppel *et al.*, 2008; Zhang *et al.*, 2011).

While these methods remain valuable, especially for large cohort studies and well-characterized conditions, they have notable limitations. Their reliance on anatomical priors and handcrafted features makes them less robust to atypical anatomy, scanner variability, or subtle pathologies. Furthermore, manual tuning and limited adaptability can further hinder scalability in heterogeneous datasets.

### 0.2.2 Recent approaches using Deep Neural Networks.

Recent advancements in artificial intelligence techniques, such as deep neural networks, have revolutionized brain MRI assessment, offering a data-driven alternative to the hand-crafted feature engineering that long dominated neuroimaging analysis. These models have shown great promise across a wide range of different tasks, enabling automated and precise identification of pathological patterns, accurate delineation of brain tissues and anatomical structures, and extraction of clinically relevant imaging biomarkers from complex MRI data (Zhao & Zhao, 2021; Ali *et al.*, 2025).

Convolutional neural networks (CNNs) first demonstrated that a model could learn hierarchical spatial features directly from raw pixels, achieving high performance in image classification tasks on benchmarks such as ImageNet (Krizhevsky, Sutskever & Hinton, 2012; He *et al.*, 2016). Building on that success, researchers adapted the same principles to volumetric MRI data, enabling robust image-level classification of neurological disorders including Alzheimer's disease (Korolev *et al.*, 2017) and distinct glioma grades (Gutta *et al.*, 2021).

Extending the success of CNNs beyond classification, researchers transformed CNN encoders into fully convolutional "U-shaped" architectures for pixel-accurate segmentation, yielding state-of-the-art performances for automatically delineate tumors, lesions, and cortical structures (Ronneberger, Fischer & Brox, 2015; Isensee *et al.*, 2021), among many others.

More recently, transformers (Vaswani *et al.*, 2017), with their self-attention mechanism, have proven effective in capturing long-range anatomical context, improving tasks that benefit from global consistency such as multi-structure parcellation and cross-slice anomaly detection (Valanarasu *et al.*, 2021; Cao *et al.*, 2022). Moreover, hybrid models that fuse CNN feature extractors with transformer blocks now combine the best of both worlds, uniting local texture sensitivity with holistic shape reasoning (Hatamizadeh *et al.*, 2022b; Zhang, Liu & Hu, 2021).

Beyond classification and segmentation, deep learning has facilitated a wide range of applications across the brain MRI pipeline. Models now support regression tasks, such as predicting brain age, offering valuable biomarkers for neurological assessment (Cole, 2017). They also improve image quality through super-resolution, denoising, and motion correction, allowing for faster and more reliable scans (Pham *et al.*, 2019; Wang *et al.*, 2023a). Furthermore, deep learning accelerates image reconstruction from raw k-space data, reducing scan time without compromising quality (Aggarwal, Mani & Jacob, 2018; Han, Sunwoo & Ye, 2019). Other important applications include predicting disease progression or survival outcomes, and performing fast, accurate registration and tractography for connectivity studies and surgical planning (Balakrishnan *et al.*, 2019).

These diverse applications highlight the broad potential of deep learning to enhance nearly every stage of brain MRI analysis, from image acquisition and reconstruction to diagnosis, prognosis, and treatment planning. Furthermore, deep learning models continually improve as more data, either annotated or unannotated, become available, making them well-suited to scale with the ever-growing volume of clinical imaging data worldwide. This adaptability, combined with their accuracy and efficiency, positions deep learning as a powerful and transformative approach in both clinical and research neuroimaging.

## 0.3 Challenges and Motivation

Despite the remarkable progress in deep learning for brain MRI analysis, several persistent challenges continue to hinder the clinical translation and generalizability of these models. These limitations not only affect their practical utility but also pose significant obstacles to the development of robust and scalable analytical tools.

One of the primary challenges in developing deep learning models for brain MRI assessment is the issue of **domain variability and distribution shifts** across imaging sources. MRI data

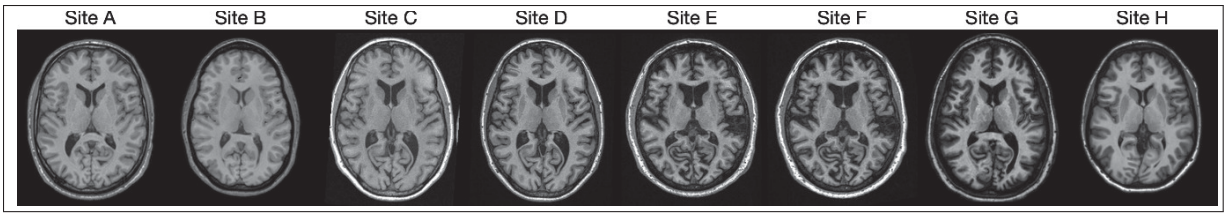



Figure 0.3 Examples of T1-weighted MRI scans from eight different sites, illustrating the variability in image contrast across acquisition settings
Adapted from Zuo *et al.* (2021)

are inherently heterogeneous due to differences in scanner hardware, magnetic field strengths, acquisition protocols, spatial resolutions, and clinical environments, all of which contribute to substantial variations in image intensity, contrast, and noise characteristics. Examples of MRI variability across 8 different sites are visualized in Figure 0.3. Consequently, deep learning models trained on data from a single source often fail to generalize effectively to datasets from other institutions or scanning devices, as these models frequently become overly sensitive to subtle, domain-specific imaging features (Karani *et al.*, 2018; Glocker *et al.*, 2019). A straightforward solution that may come to mind is to retrain the model on each new or target domain. However, this approach poses several practical limitations: (i) it requires access to labeled data for every new domain, (ii) it assumes prior knowledge of all domains the model may encounter, and (iii) it necessitates costly re-training of the model—potentially involving all previously seen data—each time a new domain is introduced. Moreover, the scarcity of diverse, multi-institutional datasets exacerbates the problem, as publicly available data tend to be limited to specific scanners and clinical settings, lacking sufficient representational diversity. Addressing domain variability and improving model robustness across distribution shifts remain critical for the broader adoption and clinical translation of deep learning-based MRI analysis methods (Bento *et al.*, 2022).

Another significant challenge is the **limited availability of high-quality annotated datasets**. Supervised learning approaches, which dominate current methodologies, require extensive

labeled data, particularly at the voxel level for tasks such as tissue or lesion segmentation. However, manual annotation is inherently costly and time-consuming, as it demands specialized expertise from radiologists or neurologists to accurately identify and annotate pathological structures such as lesions, tumors, and developmental anomalies (Litjens *et al.*, 2017). Additionally, the manual annotation process introduces inter-annotator variability, as experts may interpret imaging features differently based on their training and experience (Menze *et al.*, 2014; Yang *et al.*, 2023). This variability reduces the consistency and reliability of annotated datasets, further complicating model training and validation. Consequently, large-scale, diverse, and consistently labeled MRI datasets remain scarce, limiting the effectiveness and generalizability of supervised deep learning approaches in clinical neuroimaging applications (Esteva *et al.*, 2019).

To address the high cost of expert annotation required for fully-supervised learning, researchers have explored label-efficient learning paradigms as a solution. In semi-supervised setups, a small core of meticulously labeled scans is mixed with hundreds or thousands of unlabelled ones. Algorithms such as Transformation Consistency (Bortsova *et al.*, 2019), Mean Teacher (Tarvainen & Valpola, 2017), and FixMatch (Sohn *et al.*, 2020) encourage the network to make stable predictions across random augmentations, yielding promising segmentation performance at a fraction of the annotation cost, although still not matching the accuracy of fully supervised baselines. Weakly supervised approaches relax the requirements even further: models learn from coarse cues such as image-level disease tags, a sparse set of slice labels, or a handful of point clicks, yet still manage to localize tumors and delineate lesions with clinically useful precision with far less radiologist time (Wu *et al.*, 2019; Campanella *et al.*, 2019)

However, even with label-efficient learning paradigms, the inherent heterogeneity of brain injuries and neurological pathologies further complicates the development of accurate and robust MRI-based diagnostic models. Indeed, brain pathologies exhibit substantial variability in their anatomical locations, lesion sizes, morphological characteristics, and progression

patterns meaning that even well-annotated datasets may under-represent rare or atypical presentations (Bakas *et al.*, 2018). This extensive spectrum of potential abnormalities poses significant challenges, even for fully-supervised approaches, which typically rely on well-defined pathological and neatly defined categories and sufficient labeled examples for each specific condition. Consequently, supervised methods often fail to detect pathologies that fall outside their predefined training scope, severely limiting their clinical utility.

Finally, subtle brain anomalies that are **undetectable through conventional structural MRI analysis** present an additional critical challenge. Such anomalies are particularly evident in neonatal populations, where brain development issues, such as intrauterine growth restriction (IUGR) or subtle neurodevelopmental delays, often manifest in subtle disruptions of normal brain maturation processes rather than clear structural abnormalities. In such cases, standard structural MRI analysis falls short, and there is a need for analytical approaches that go beyond visual inspection, specifically, methods capable of extracting developmental biomarkers that quantify deviations from normative growth trajectories (Sun *et al.*, 2024; Fleiss *et al.*, 2019).

Taken together, the abovementioned issues underscore the urgent need for more resilient, data-efficient, and interpretable approaches to brain MRI analysis. Addressing these challenges is crucial to advancing the field of neuroimaging, enhancing diagnostic capabilities, and improving patient care through more precise, robust, and universally applicable analytical tools.

## 0.4 Generative Models for Brain MRI Assessment: Rationale and Potential

*"What I cannot create, I do not understand."* — *Richard P. Feynman*

Generative models have emerged as a powerful paradigm in machine learning, grounded in the idea that true understanding of data stems from the ability to simulate or recreate it. Unlike discriminative models, which focus on mapping inputs to outputs, generative models aim to

learn the underlying probability distribution of the data itself. This fundamental distinction gives generative models unique capabilities that are particularly relevant for complex domains such as brain MRI analysis.

One of the key advantages of generative models lies in their ability to learn from unlabeled data. In medical imaging, acquiring high-quality annotations is often prohibitively expensive and time-consuming, requiring expert knowledge and considerable manual effort, in addition to be prone to annotator variability. Generative models, such as variational autoencoders (VAEs) (Kingma, 2013), generative adversarial networks (GANs) (Goodfellow *et al.*, 2014), and diffusion models (Ho, Jain & Abbeel, 2020b), can leverage large collections of unlabeled scans to approximate the distribution of relevant features (e.g., anatomical structures), thereby reducing dependence on annotated datasets and enabling broader scalability.

Another important strength is their capacity to capture and simulate variability in brain anatomy and appearance. This is particularly useful in neuroimaging, where inter-subject differences, developmental stages, and biological diversity contribute to substantial variation even among healthy individuals. Additionally, generative models naturally support interpretability, which is crucial in clinical settings. Their generative nature offers intuitive mechanisms for visualizing model behavior, for example, by observing how reconstructions deviate from real images or how latent space interpolations translate into anatomical changes.

In summary, generative models offer a flexible and scalable complement to traditional discriminative approaches in brain MRI analysis. Their ability to learn from unlabeled data, represent structured variability, and provide interpretable outputs makes them well-suited to address key challenges in medical imaging, particularly in data-constrained and high-variance settings.

## 0.5 Proposed Solutions and Research Objectives

Addressing the challenges outlined in the previous sections requires the development of innovative methodologies that ensure both generalizability and robustness across diverse clinical settings and a wide range of brain pathologies. As discussed in the previous section, generative models provide a principled framework well-suited to these demands, particularly in scenarios where labeled data is limited, anatomical variability is high, and interpretability is critical. Leveraging these advantages, this thesis mostly focuses on approaches based on generative models for advancing brain MRI analysis.

For learning-based MRI analysis methods to be effectively translated into clinical practice, the proposed methods must: (i) maintain consistent performance across different scanners and imaging sites, (ii) operate without relying on exhaustive voxel-wise annotations while accurately detecting a broad spectrum of abnormalities, and (iii) identify subtle maturational deviations that often elude standard visual assessment. This chapter introduces the proposed solutions and research objectives pursued in this thesis, which center on three key areas: Unsupervised MRI harmonization, unsupervised anomaly detection, and neonatal brain age estimation. Each of these objectives is discussed in detail in the following sections, outlining their motivation and rationale, the methodologies developed, and their significance in advancing clinically applicable brain MRI analysis.

### 0.5.1 MRI Harmonization

**"Unsupervised MRI harmonization"** refers to the process of aligning medical images from different scanners, protocols, or sites without relying on labeled or paired data. This task is critical in brain MRI analysis, where domain shifts introduced by variations in acquisition hardware and procedures often degrade the performance of machine learning models (Abbasi *et al.*, 2024). By standardizing image appearance across sources, harmonization aims at improving consistency

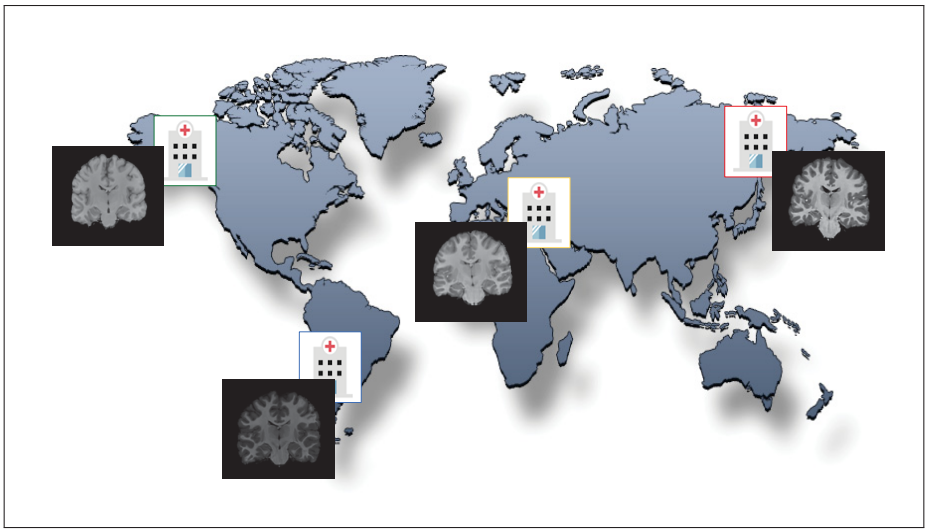


Figure 0.4 MRI variability across different scanners

in quantitative imaging metrics and allows the integration of multi-site data. Importantly, unsupervised methods address this challenge without the need for manual annotations, paired data or domain-specific supervision, making them highly scalable for real-world applications.

MRI Harmonization directly addresses the generalization limitations of deep learning models trained on domain-specific MRI data. Unsupervised harmonization mitigates the distribution shift problem by learning to transform images from disparate sources into a common, domain-invariant space. This process neutralizes scanner- or site-specific artifacts while preserving essential anatomical and pathological features. By doing so, it ensures that diagnostic models trained on harmonized data maintain performance across unseen datasets and clinical environments.

### 0.5.2 Unsupervised Anomaly Detection

**"Unsupervised anomaly detection"** in brain MRI is a transformative approach that enables the identification of pathological deviations without requiring extensive annotated training data. Unlike traditional supervised techniques, which depend on curated examples of specific anomalies, unsupervised methods operate by modeling the distribution of healthy brain scans

and detecting deviation from this learned normality. This strategy is particularly advantageous in medical imaging, where annotated datasets are often scarce, expensive to produce, and biased toward common pathologies. Unsupervised detection offers broader generalization and the ability to flag rare or previously unseen abnormalities, making it a powerful tool for scalable and accessible diagnostic support.

This approach directly addresses key clinical and methodological challenges in brain MRI analysis, such as data imbalance, rare disease detection, and reduces reliance on manual annotations, which are often subject to inter-observer variability. Furthermore, machine learning models trained solely on annotated datasets often fail to generalize to new clinical scenarios, especially when encountering anomalies not present during training. Unsupervised methods overcome this limitation by learning the manifold of normal anatomy and identifying outliers based on deviation rather than category. This makes them inherently robust to novel pathologies, protocol shifts, and subtle structural abnormalities, challenges that frequently undermine supervised approaches in real-world deployments.

### 0.5.3 Neonatal Brain Age Estimation

To ensure a comprehensive assessment of brain MRI, we explore **"Neonatal brain age estimation"** which is a method designed to detect subtle developmental delays in infants by predicting brain maturity directly from MRI scans. Traditional structural MRI, while effective for identifying gross anatomical abnormalities, often fails to capture early neurodevelopmental disruptions such as intrauterine growth restriction (IUGR) or delayed myelination. These conditions may not produce clear structural changes during the neonatal period, making early diagnosis difficult. Brain age estimation offers an alternative by quantifying developmental progress and identifying deviations from expected maturation trajectories.

Neonatal brain age estimation addresses a critical limitation in neonatal neuroimaging: the inability of conventional MRI to detect early-stage or regionally confined developmental abnormalities. By modeling the normal trajectory of brain growth across regions, brain age estimation enables the identification of infants whose brain development lags behind their chronological age. This discrepancy serves as a biomarker for subtle or emerging neurodevelopmental conditions that might otherwise go unnoticed. Importantly, early detection through such predictive modeling allows for timely interventions during periods of heightened neuroplasticity, potentially improving long-term cognitive and functional outcomes.

## 0.6 Contributions

The core contributions of this thesis are:

- In Chapter 2, we propose a novel harmonization framework, "Harmonizing Flows," which leverages the power of normalizing flows, a class of invertible and expressive generative models, for source-free, unsupervised MRI harmonization. Unlike conventional harmonization strategies that require paired samples or explicit domain labels, our method learns to model the underlying distribution of each source domain independently and maps them into a shared latent representation. The invertibility of normalizing flows ensures that no information is lost during transformation, preserving fine-grained anatomical detail essential for downstream clinical tasks such as anomaly detection and brain age estimation. Our empirical results demonstrate the effectiveness and robustness of the proposed approach across diverse datasets and acquisition settings. We show that predictive models trained on harmonized data exhibit significantly improved cross-domain generalization, maintaining high diagnostic accuracy even when evaluated on previously unseen sites. Furthermore, the preservation of medically relevant features was quantitatively and qualitatively validated, confirming the model's ability to balance harmonization and diagnostic integrity. These contributions highlight the potential of "Harmonizing Flows" to become a practical tool

for large-scale, multi-site neuroimaging studies and a foundation for deploying robust AI systems in heterogeneous clinical settings.

**Related publication:**

***Farzad Beizaee**, Christian Desrosiers, Gregory A. Lodygensky, and Jose Dolz. "Harmonizing Flows: Unsupervised MR harmonization based on normalizing flows" Presented at International Conference of Information Processing in Medical Imaging (IPMI), 2023.*

***Farzad Beizaee**, Gregory A. Lodygensky, Chris L. Adamson, Deanne K. Thompson, Jeanie LY Cheong, Alicia J. Spittle, Peter J. Anderson, Christian Desrosiers, and Jose Dolz. "Harmonizing Flows: Leveraging normalizing flows for unsupervised and source-free MRI harmonization." Accepted at Journal of Medical Image Analysis, 2025.*

- In Chapters 3, 4, and 5, we introduce a progression of generative models for unsupervised brain anomaly detection, beginning with DeCo-Diff, a reformulated diffusion model originally tested on highly variable industrial datasets. Demonstrating its effectiveness in a domain more complex than typical medical imaging provided strong evidence of the method's generalization capacity. Building on this foundation, we developed MAD-AD, Masked Diffusion for Anomaly Detection, an architecture that strategically applies masking in the latent space of a diffusion model. By training exclusively on healthy scans, MAD-AD learns a detailed internal representation of normal brain structures and uses targeted reverse diffusion to reconstruct anomalies with high precision, isolating them from the normal context.

  To further improve robustness and consistency, we proposed REFLECT, a rectified flow-based model that replaces the stochasticity of diffusion models with deterministic transport dynamics. REFLECT enhances the interpretability, stability, and repeatability of anomaly detection outcomes, addressing a key clinical barrier to adoption: the unpredictability of generative models. The results across multiple datasets and clinical conditions confirm that REFLECT outperforms conventional diffusion-based detectors in both localization and reconstruction fidelity. Together, these contributions form a robust and clinically relevant

pipeline for unsupervised brain anomaly detection, capable of operating in data-scarce environments while maintaining the diagnostic rigor necessary for real-world use.

**Related publications:**

***Farzad Beizaee**, Gregory A. Lodygensky, Christian Desrosiers, and Jose Dolz. "Correcting Deviations from Normality: A Reformulated Diffusion Model for Multi-Class Unsupervised Anomaly Detection." Presented at IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2025.*

***Farzad Beizaee**, Gregory Lodygensky, Christian Desrosiers, and Jose Dolz. "MAD-AD: Masked Diffusion for Unsupervised Brain Anomaly Detection." Presented at International Conference of Information Processing in Medical Imaging (IPMI), 2025.*

***Farzad Beizaee**, Sina Salimi, Ismail Ben Ayed, Gregory A. Lodygensky, Christian Desrosiers, and Jose Dolz. "REFLECT: Rectified Flows for Efficient Brain Anomaly Correction Transport." Submitted to International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2025.*

- In Chapter 6, in order to complete the final piece of the puzzle in my thesis on brain MRI assessment, we have explored neonatal brain age estimation with less focus on generative modeling aspect. we proposed a simple, yet effective deep learning-based framework titled "Determining regional brain growth in premature and mature infants in relation to age at MRI using deep neural networks". This method is trained on a cohort of healthy neonatal brain MRIs and learns to predict brain age based on regional imaging features. By comparing the predicted brain age to the actual age at scan time, we generate individualized developmental maturity profiles. Infants exhibiting significant negative age gaps, where the brain appears developmentally younger than expected, can be flagged for further clinical evaluation, even when standard imaging assessments appear normal.

  Our model demonstrates strong predictive accuracy and clinical utility, validated across diverse neonatal populations and imaging conditions. Beyond global age prediction, the

model provides fine-grained, region-specific assessments of brain growth, enabling the identification of localized developmental delays in regions such as the hippocampus or thalamus. This level of detail supports precision medicine approaches and longitudinal monitoring, offering a powerful tool for both early diagnosis and therapy evaluation. By integrating predictive modeling with clinical neuroimaging, our contribution enhances diagnostic sensitivity and paves the way for more personalized and proactive neonatal care.

**Related publication:**

***Farzad Beizaee**, Michele Bona, Christian Desrosiers, Jose Dolz, and Gregory A. Lodygensky. "Determining regional brain growth in premature and mature infants in relation to age at MRI using deep neural networks." Accepted in Journal of Scientific Reports, 2023.*

## 0.7 Integration and Broader Impact

The solutions presented in this thesis are deeply interrelated, collectively addressing key limitations in brain MRI analysis. MRI harmonization ensures consistent and comparable imaging data across scanners and acquisition protocols, forming a reliable foundation for downstream analysis. Unsupervised anomaly detection introduces scalable, label-efficient tools for identifying diverse pathologies, while neonatal brain age estimation enables the early detection of subtle neurodevelopmental delays that may not manifest in visual inspection of structural imaging alone.

Together, these methods form a cohesive framework for improving the accessibility, reliability, and interpretability of brain MRI analysis. They contribute to a broader shift in medical AI: moving beyond narrowly optimized models trained on idealized datasets toward clinically grounded robust systems capable of generalizing across diverse real-world settings. Each method has been validated across different datasets, use cases, and patient populations, reflecting a commitment to generalization and translational impact.

In conclusion, the integration of generative-model-based methods into brain MRI assessment holds transformative potential for clinical neuroimaging, healthcare delivery, and patient care globally. The advancements presented in this thesis contribute to overcoming critical barriers to clinical adoption, setting a foundation for future research directions and practical applications that are both scientifically innovative and socially impactful.

# CHAPTER 1

# LITERATURE REVIEW

## 1.1 Preliminaries: Generative models

Generative models play an increasingly central role in medical image analysis by enabling models to learn the distribution of data and generate plausible samples from it. In the context of brain MRI, this capability supports a range of different tasks including realistic data synthesis (Pinaya *et al.*, 2022b), anomaly detection (Behrendt *et al.*, 2024b), image enhancement (Wu *et al.*, 2023b), modality translation (Hiasa *et al.*, 2018), and image reconstruction (Song *et al.*, 2022). The ability of generative models to capture complex distributions of structural and developmental brain features makes them especially valuable in clinical settings where labeled data is limited, population diversity is high, and pathological variations are subtle.

Generative models encompass a diverse family of techniques, including variational autoencoders (VAEs) (Kingma & Welling, 2014), generative adversarial networks (GANs) (Goodfellow *et al.*, 2014), normalizing flows (Rezende & Mohamed, 2015), score-based diffusion models (Song *et al.*, 2021c; Ho *et al.*, 2020b), and recently, the family of continuous normalizing flows (CNFs) (Chen *et al.*, 2018; Liu, Gong & Liu, 2023a; Lipman *et al.*, 2022) each offering complementary strengths in flexibility, likelihood tractability, or sample fidelity. In particular, this thesis leverages three likelihood-based generative models, normalizing flows, diffusion models, and a recent hybrid approach called rectified flows that aim to combine their strengths. The following sections provide a conceptual overview of these generative models.

### 1.1.1 Normalizing Flows

Normalizing flows are a class of generative models that construct complex probability distributions by applying a sequence of invertible transformations to a simple base distribution. Formally, let $x \in \mathbb{R}^D$ denote an observed data sample and $z_0 \sim p_0(z_0)$ denote a sample drawn from a base distribution $p_0$ (often a standard Gaussian). By applying $K$ successive bijective

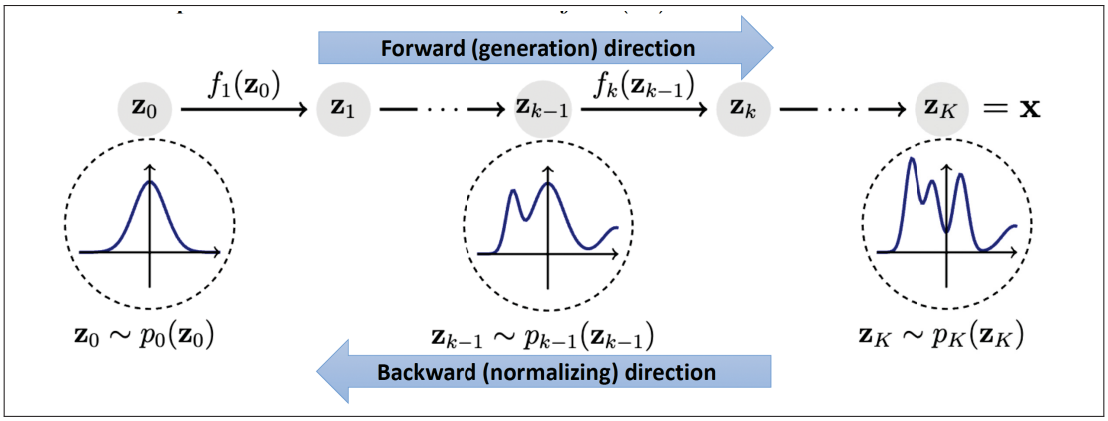



Figure 1.1 Normalizing flow distribution transformation. The forward process converts a simple distribution into a complex one via sequential transformations. This process is fully invertible, enabling exact likelihood estimation via the backward pass
Taken from Shen & Shen (2023)

functions ($z_k = f_k(z_{k-1}; \theta_k)$) parameterized by $\theta_k$ where $k = 1, \ldots, K$, so that $x = z_K$, and leveraging the change-of-variables formula, $x$ follows a distribution whose density can be evaluated using:

$$\log p_\theta(x) = \log p_Z(z_0) - \sum_{k=1}^{K} \log\left|\det J_{f_k}(z_{k-1})\right|, \tag{1.1}$$

where $J_{f_k}(z_{k-1}) = \frac{\partial f_k(z_{k-1};\theta_k)}{\partial z_{k-1}}$ is the Jacobian of the $k^{th}$ transformation. This yields an explicit likelihood model where the density $p_\theta(x)$ can be computed exactly, and new samples can be generated by sampling $z_0 \sim p_0(z)$ and applying the sequence of transformations $x = f_K \circ \cdots \circ f_1(z_0)$. Also, training is typically performed via maximum likelihood estimation by maximizing $\log p_\theta(x)$ over the dataset.

In practice, each transformation function ($f_k$) can be implemented as a neural network, as long as it is bijective and its Jacobian log-determinant remains tractable and efficiently computable. Early seminal works like NICE (Nonlinear Independent Components Estimation) (Dinh, Krueger & Bengio, 2015) introduced coupling layers as a practical invertible transform,

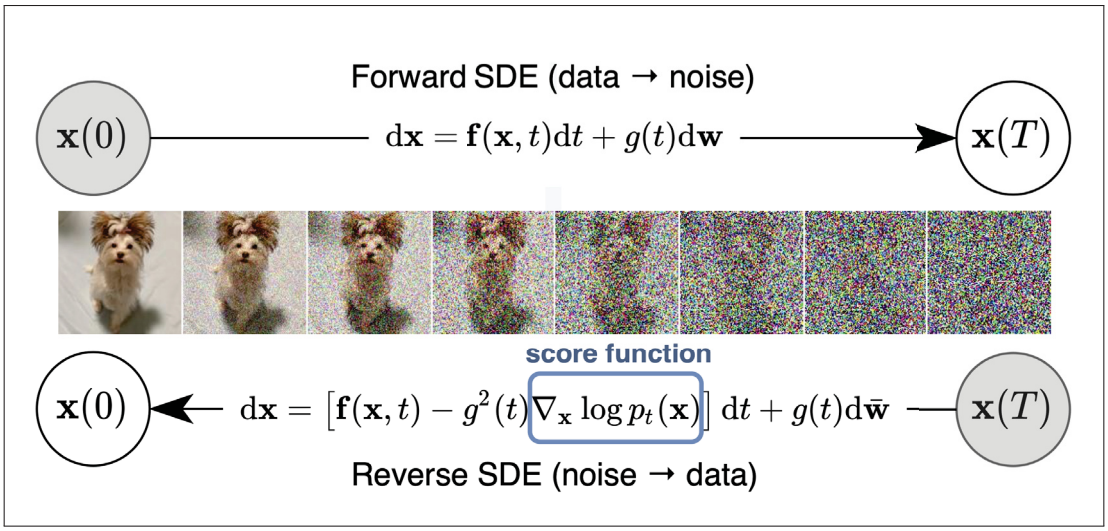



Figure 1.2 Score-based diffusion model forward and Reverse SDE transformations
Taken from Song *et al.* (2021c)

followed by RealNVP (Real-valued Non-Volume Preserving flow) (Dinh *et al.*, 2017), which added scale transforms, and Glow (Kingma & Dhariwal, 2018), which further improved architecture (e.g. using invertible $1 \times 1$ convolutions for mixing channels).

By offering tractable densities and stable training under a maximum-likelihood objective, normalizing flows avoid the adversarial instability and mode collapse often encountered in GANs (Goodfellow *et al.*, 2014), while also bypassing the variational approximations of VAEs (Rezende & Mohamed, 2015) that can lead to overly smooth or blurry samples. As a result, flow-based models not only deliver exact likelihood evaluations but also generate sharp, high-fidelity samples without the need for auxiliary inference networks.

### 1.1.2 Diffusion Models

Diffusion models are a class of generative models that achieve state-of-the-art results in image synthesis by learning to reverse a gradual noising process applied to data. The core idea, introduced in early work by Sohl-Dickstein *et al.* (2015) and later refined by others (Ho *et al.*, 2020b; Song, Meng & Ermon, 2021a; Song *et al.*, 2021c), is to start with a complex data

distribution (e.g., MRI images) and diffuse it into random noise through a sequence of small perturbations, then learn how to invert that sequence.

Let $x_0 \in \mathbb{R}^D$ denote a data sample (e.g., a brain MRI), and define a forward diffusion process $\{x_t\}_{t=0}^T$ where noise is added over $T$ discrete time steps via a Markov chain $q(x_t \mid x_{t-1})$. Typically, this forward process is defined as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t\, \mathbf{I}), \tag{1.2}$$

where $\beta_t \in (0, 1)$ is a small variance schedule controlling the noise magnitude at step $t$. Under this process, the marginal posterior $q(x_t \mid x_0)$ remains Gaussian and can be written in closed form:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\, \mathbf{I}), \tag{1.3}$$

with $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. As $t \to T$, $x_T \sim \mathcal{N}(0, \mathbf{I})$ becomes nearly pure noise.

The generative process attempts to reverse this diffusion. In DDPM, a neural network $\epsilon_\theta(x_t, t)$ is trained to predict the added noise, and the reverse transitions are defined as:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \tag{1.4}$$

where $\mu_\theta$ and $\Sigma_\theta$ are derived functions of the noise prediction $\epsilon_\theta$. Training minimizes a reweighted variational bound, which under certain simplifications reduces to a denoising objective:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2\right]. \tag{1.5}$$

An alternative, more general formulation is provided by score-based diffusion models (Song *et al.*, 2021c), which learn the score function $\nabla_{x_t} \log q(x_t)$ at each noise level using score matching. Sampling is then performed by numerically solving a corresponding stochastic differential equation (SDE).

Notably, these two views were later shown to be equivalent in fundamental ways and demonstrated that the objective of DDPM is essentially a weighted form of the score-matching objective (Song *et al.*, 2021b). Thus "diffusion models" broadly refer to this family of approaches that generate data by gradually denoising random noise, whether implemented as discrete time-step chains (DDPM) or continuous SDEs (score-based generative modeling). Crucially, diffusion models do not require the transformations to be bijective in the same way flows do; instead, they rely on running a generative Markov chain. The absence of invertibility constraints and the strong theoretical foundation (each step can be seen as modeling a well-defined conditional or score) allow diffusion models to represent extremely complex distributions given sufficient steps.

### 1.1.3 Rectified Flows

Rectified flows are a recent advancement in continuous generative modeling that aim to combine the exact likelihood and sampling efficiency of normalizing flows with the sample quality and flexibility of diffusion models (Liu *et al.*, 2023a). Rectified flows define a deterministic generative process via an ordinary differential equation (ODE) that transforms a base distribution $z \sim p_0(z)$ (typically standard Gaussian) into a complex data distribution $x \sim p_\theta(x)$, but with a key structural constraint: the transport trajectories should be as straight as possible in data space.

Let $x(t) \in \mathbb{R}^D$ denote a trajectory evolving over time $t \in [0, 1]$, governed by the ODE:

$$\frac{dx}{dt} = v_\theta(t, x(t)), \tag{1.6}$$

with initial condition $x(0) = z \sim p_0(z)$. The terminal state $x(1)$ represents a sample from the target distribution. Unlike generic continuous normalizing flows (CNFs) (Chen *et al.*, 2018), where trajectories can be arbitrarily curved, rectified flows encourage $x(t)$ to follow approximately linear paths from $x(0)$ to $x(1)$. This rectification constraint is motivated by optimal transport theory, a straight-line path (in expectation) defines the shortest (i.e., straightest) transport path.

To enforce this rectification constraint, the training objective is designed to align the evolving distribution $p_\theta(x(t))$ at intermediate times $t \in (0, 1)$ with a linear interpolation between the base and target distributions. Specifically, the target distribution at time $t$ is defined as:

$$p_t(x) = (1 - t)\, p_0(x) + t\, p_{\text{data}}(x), \tag{1.7}$$

which encourages the generative trajectory to follow a straight path in distribution space from $p_0$ to $p_{\text{data}}$.

Rather than simulating full stochastic processes or score estimation as in diffusion models, rectified flows directly learn the ODE vector field $v_\theta(t, x)$ by minimizing a time-dependent distributional discrepancy between $p_\theta(x(t))$ and $p_t(x)$ over randomly sampled time steps.

Crucially, the approximate linearity of trajectories enables highly efficient sampling. The learned ODE can be discretized using a small number of integration steps ($N$), and in extreme cases, a single forward Euler step (i.e., $x(1) \approx x(0) + v_\theta(0, x(0))$) suffices to generate high-quality samples.

To further refine the learned transport trajectories, they have also proposed a recursive procedure called *Reflow*. After training an initial rectified flow model, its generative distribution is treated as a new base distribution $p_0^{(1)}$, and the interpolation target is updated to $p_t^{(1)}(x) = (1 - t)\, p_0^{(1)}(x) + t\, p_{\text{data}}(x)$. This process can be repeated over multiple stages, progressively straightening the transport trajectories and improving sample fidelity. Reflow thus acts as a form of iterative refinement, encouraging convergence toward the optimal transport path and reducing the number of integration steps required for high-quality generation.

Therefore, rectified flows provide a compelling generative modeling framework that combines the continuous and interpretable dynamics of ODE-based methods with high sampling efficiency. By avoiding hundreds of stochastic or deterministic steps which are typically required by score-based diffusion models, and enabling high-fidelity generation with minimal integration steps, they are

particularly well-suited for high-dimensional applications like brain MRI synthesis, where both generation quality and computational efficiency are critical.

## 1.2 MRI harmonization

Multi-center MRI studies face significant domain shifts due to variations in scanner hardware, vendor-specific configuration, magnetic field strength, and imaging protocols. These differences lead to inconsistencies in image intensity distributions, contrast, and spatial resolution, leading to non-biological variability across datasets. Such variability not only undermines the reliability of multi-cohort studies but also poses a challenge to the reproducibility of statistical analyses and biomarker discovery. Moreover, it hampers the generalizability and robustness of downstream learning-based models, which often rely on consistent data distributions to perform accurately. To address these challenges, MRI harmonization techniques have been developed to reduce inter-scanner and inter-protocol discrepancies while preserving anatomically meaningful information. Broadly, these methods fall into two categories: (i) traditional statistical harmonization techniques that adjust or standardize imaging intensities using predefined transformations or statistical models, and (ii) learning-based harmonization methods that leverage deep learning techniques to learn mappings between scanners or domains. We review each category in detail below.

### 1.2.1 Traditional Statistical Methods

Early approaches to MRI harmonization relied on statistical normalization and intensity standardization. Simple methods include z-score normalization (Nyúl & Udupa, 1999) and histogram matching (Nyúl, Udupa & Zhang, 2000), which adjust image intensities to a reference distribution. Another intensity-based method is WhiteStripe normalization (Shinohara *et al.*, 2014). WhiteStripe normalization focuses on a specific landmark: the normal-appearing white matter intensity. It automatically finds a narrow intensity range in each scan that corresponds to peak white matter signal (the “white stripe” in the histogram) and then normalizes the image such that the median white matter intensity is set to a fixed value.

A more advanced and widely used statistical method is ComBat, originally developed for batch-effect correction in gene expression data (Johnson, Li & Rabinovic, 2007) and later adapted for imaging. In neuroimaging, ComBat can be applied either to derived measures, e.g., cortical thickness (Fortin *et al.*, 2018), diffusion MRI metrics (Fortin *et al.*, 2017), or voxel intensities (Pomponio *et al.*, 2020). ComBat performs location-scale adjustment of intensities across scanners or sites, using an empirical Bayes framework to remove scanner-specific biases while retaining biological variability.

These statistical methods established the importance of harmonization in multi-scanner studies and serve as a baseline for more recent machine learning approaches. These approaches generally do not require large training datasets and are model-agnostic, making them simple and widely applicable. However, they often assume a global intensity mapping and may struggle with complex, nonlinear differences between imaging protocols. Moreover, they often need to be re-calibrated when a new site is introduced, limiting scalability.

### 1.2.2 Learning-Based Harmonization Approaches

The advent of deep learning has brought a significant paradigm shift in MRI harmonization, moving from handcrafted transformations and statistical adjustments toward data-driven frameworks capable of modeling complex, non-linear relationships between imaging domains (Abbasi *et al.*, 2024). Learning-based harmonization methods aim to reduce scanner- and protocol-induced variability by synthesizing images that emulate a predefined reference domain, such as a specific scanner model, magnetic field strength, or acquisition protocol, while faithfully preserving subject-specific anatomical structures.

**Supervised CNN-Based Harmonization (Paired Data)** One important subclass of learning-based methods uses supervised learning with paired data from multiple scanners. A common scenario is a traveling subject study, where the same individuals are scanned on multiple scanners or protocols, providing direct correspondences. DeepHarmony (Dewey *et al.*, 2019) is an example of such methods that uses a U-Net convolutional neural network to learn a voxel-wise

intensity mapping between MRI acquisitions. Trained on paired scans from different scanners, DeepHarmony can adjust contrast differences across scanner changes in an end-to-end manner. Tian *et al.* (2022) proposed a method that uses a deep learning framework trained on traveling subject data to disentangle site-specific factors from subject-specific anatomy, enabling the generation of harmonized images that preserve individual brain features while reducing inter-site variability. MISPEL is another example of supervised methods (Torbati *et al.*, 2021) which uses a two-step training (embedding alignment then intensity harmonization) that forces the latent embeddings from different scanners to be similar, so that decoding yields identical harmonized images across scanners.

Supervised approaches achieve contrast harmonization by explicitly learning the voxel-wise intensity transformation from the source domain to the target domain. They tend to perform well when plenty of paired training data (e.g., the same subjects scanned on multiple scanners) is available. However, they require the availability of overlapping scans between domains, which in practice is rare and expensive (it demands either scanning volunteers on multiple machines or using phantoms). The learned model is also specific to the scanner pair seen during training, and if a new scanner or site is introduced, additional paired data and retraining are needed, which limits the generalizability of such models.

**Unsupervised Image-to-Image Translation using Unpaired Datasets** To overcome the need for subject overlap, many works have turned to unsupervised image-to-image translation techniques. These methods learn from unpaired datasets of images from each site, often using GANs. CycleGAN-based approaches are particularly popular for unpaired image-to-image translation (Zhu *et al.*, 2017; Modanwal *et al.*, 2020). CycleGAN employs two generator-discriminator pairs and a cycle-consistency loss to learn mappings between Domain A (e.g. scanner or protocol X) and Domain B (scanner or protocol Y) without one-to-one correspondences. Gao *et al.* (2019) extends this idea by introducing a many-to-one CycleGAN for MRI intensity standardization, which maps various scanners' images to a single reference scanner's contrast.

In addition to CycleGAN-based methods, many variants have been proposed. Style transfer GANs inject style codes from reference images to harmonize without explicit domain labels. For instance, Liu *et al.* (2021) demonstrated that encoding the "style" of a reference scanner's MRI and transferring it to input images can achieve harmonization without needing to know scanner labels. Their approach, trained on multi-site data, effectively treated scanner differences as differences in image style and performed a style transfer while preserving anatomical content. Guan *et al.* (2021) proposed an attention-guided GAN for multi-site MRI that focuses the translation on contrast-relevant regions, improving brain disorder classification after harmonization.

These GAN methods can generate images that are visually similar to the target domain while retaining the subject's anatomical structures. A major advantage is that they do not require traveling subjects; instead, they leverage unpaired, unlabeled datasets from each scanner. However, a notable drawback is that these models typically require prior knowledge of the source and target domains and tend to generalize poorly to unseen scanners or acquisition protocols. IGUANe (Roca *et al.*, 2025), a 3D CycleGAN variant, tries to address this by incorporating an attention mechanism and a unified generator designed to harmonize images across multiple sites and even to previously unseen domains. Nevertheless, the inherent instability of adversarial training and the potential for hallucinating artificial structures, especially when there is a population mismatch between source and target domains (e.g., healthy vs. pathological), remain critical challenges for clinical deployment.

**Disentangled Representation Learning for Harmonization.** Another line of research seeks to disentangle scanner/protocol effects in a latent representation, often using autoencoders or VAEs. The idea is to learn a disentangled representation that separates imaging data into "content" (anatomy) and "style" (scanner-specific appearance), instead of directly translating images. The network can then recombine content with a desired style to generate a harmonized image. For example, Dewey *et al.* (2020) introduced a VAE-based model that learns a disentangled latent space for cross-site MRI harmonization. By enforcing that one part of the latent vector captures only site-specific variations, they can then adjust or remove that part before reconstructing the image, effectively standardizing the output across sites.

Zuo *et al.* (2021) extended this idea with an information bottleneck VAE, achieving unsupervised harmonization by penalizing mutual information between latent factors and domain labels. HACA3 (Zuo *et al.*, 2023) introduces a novel framework that disentangles MRI images into three components: anatomical structure, contrast information, and imaging artifacts. This separation allows for more accurate harmonization across different imaging sites and protocols. HACA3 employs an attention mechanism to effectively fuse these components, ensuring that the harmonized images maintain anatomical integrity while adapting to the desired contrast and minimizing artifacts.

These methods often rely on paired multi-modal MRIs for training, thereby restricting their applicability, particularly in single-modality scenarios. Two recent feature-disentangle-based methods, i.e. ImUnity (Cackowski *et al.*, 2023) and DLEST (Wu *et al.*, 2025) tried to address this problem. ImUnity, a 2.5D VAE-GAN hybrid designed for multi-center MRI harmonization, combines a variational autoencoder with a GAN and a domain confusion module to learn a domain-invariant latent space. DLEST achieved this by disentangling anatomical content and scanner-specific style within a low-dimensional latent space, utilizing an energy-based model for style translation.

Feature disentanglement-based harmonization models offer several key advantages. By explicitly separating anatomical content from scanner-specific style, they are better able to preserve structures during harmonization. Additionally, once disentangled, the anatomical representation can be flexibly recombined with various scanner styles, enabling harmonization to multiple target domains using a single model. However, ensuring a perfect split between content and style representation is challenging, and some scanner effects might entangle with anatomy. Therefore, While disentanglement models have better generalization potential than direct image-to-image translation methods, they still require diverse and representative training data to be robust to unseen domains. Also, training these models typically involves complex architectures and multiple competing loss functions which can make optimization unstable and model behavior difficult to interpret. As a result, their performance and generalizability often depend heavily on careful tuning and validation.

**Source-free Blind MRI Harmonization.** A recent and highly practical line of research focuses on source-free blind harmonization, where the harmonization model is trained solely on source domain data without access to any prior knowledge of target domain data, hence the term "blind." These models are also considered "source-free" because, once trained, they no longer require access to the source dataset during inference. Instead, they are expected to generalize to unseen target domains without the need for retraining. Our proposed work in this thesis falls into this category and it was one of the pioneering works in this direction.

Another concurrent work is BlindHarmony (Jeong *et al.*, 2023), which introduces a flow-based generative model trained exclusively on source domain images. At test time, the model harmonizes unseen target images by optimizing a harmonized image that closely matches the target input while remaining within the source distribution. BlindHarmony has shown strong performance on both synthetic and real-world multi-site datasets without requiring multi-centric or paired data. Building on this idea, BlindHarmonyDiff (Jeong *et al.*, 2025) addresses 3D MRI harmonization challenges, such as inter-slice inconsistency and large domain shifts, by incorporating a rectified flow model trained to reconstruct images from structural edge maps. It introduces a multi-stride patch training strategy and a refinement module to produce anatomically consistent and artifact-free harmonized volumes.

Source-free blind harmonization methods offer several advantages: they do not require access to source data at inference, prior knowledge on target data, traveling subjects, multi-modal acquisitions, or labeled annotations, making them suitable for scalable and privacy-preserving deployments. They generalize well to unseen scanners compared to traditional image-to-image translation models and feature-disentangle-based methods and they often preserve subject-specific anatomy more reliably due to their self-supervised or distribution-aware training. Nonetheless, ensuring clinical reliability and anatomical fidelity remains essential for their broader adoption in practice.

In summary, existing brain MRI harmonization methods have made progress in reducing site-specific variability, yet many remain limited by their reliance on a target task, traveling

subjects, or paired modalities. Moreover, most approaches do not generalize well to unseen domains or preserve subtle anatomical details critical for downstream analyses. To address these gaps, Chapter 2 introduces an unsupervised harmonization framework based on normalizing flows, which alleviates the aforementioned constraints of previous methods, while demonstrating superior performance in harmonizing multi-site brain MRI data.

## 1.3 Unsupervised Anomaly Detection

Detecting anomalies in medical imaging involves identifying regions that deviate from normative anatomy, such as tumors, lesions, or malformations, which may indicate underlying pathology. Conventional computer-aided diagnosis systems have traditionally relied on supervised learning, requiring large volumes of annotated data for each specific disease type. However, such annotations are expensive to obtain, often subject to inter-rater variability, and inherently limited in scope: models trained on known conditions tend to perform poorly when encountering rare or previously unseen abnormalities.

Unsupervised anomaly detection (UAD) offers a promising alternative by learning the distribution of healthy anatomy alone and identifying deviations as potential anomalies, without relying on any pathological labeled data. In this paradigm, a model is trained exclusively on normal brain MRIs and subsequently used to flag regions in new scans that diverge from the learned concept of "normal." This formulation not only alleviates the dependency on labeled pathological data but also enables the detection of a broader and potentially open set of anomalies. UAD is particularly well-suited to brain MRI, where the diversity of pathologies, subtlety of certain conditions, and scarcity of labeled data create challenges for supervised methods.

Existing UAD methods in brain MRI can be broadly categorized into three families: (i) *embedding-based methods*, which identify outliers in learned feature spaces; (ii) *synthetic-based methods*, which use artificially generated anomalies to train surrogate detectors; and (iii) *reconstruction-based methods*, which rely on models trained to reproduce healthy anatomy and flag deviations in reconstruction as anomalies. The following sections review representative

works in each category, discuss their core assumptions and methodologies, and highlight their respective strengths and limitations.

### 1.3.1 Embedding-Based UAD Methods

Embedding-based anomaly detection methods aim to learn compact and semantically meaningful representations of normal brain MRIs in a latent feature space. Anomalies are then identified as outlier, data points that deviate from the learned distribution of normal embeddings. Several conceptual paradigms have emerged within this framework, differing in how they define the embedding space and how outliers are scored.

A classical approach is one-class learning using Deep Support Vector Data Description (Deep SVDD) (Ruff *et al.*, 2018), where the goal is to map all normal embeddings into a minimal hypersphere in feature space. At inference, samples lying far from this compact center are flagged as anomalous. Patch-based extensions such as PatchSVDD (Yi & Yoon, 2020) apply this principle at the patch level, enabling spatial anomaly maps by treating each image patch independently.

Recent advances leverage self-supervised and contrastive learning to shape the embedding space more effectively. PatchCL-AE (Lu *et al.*, 2024) incorporates contrastive objectives at the patch level within an autoencoding framework, enforcing local consistency among features extracted from healthy tissue. Similarly, CRADL (Lüth *et al.*, 2023) uses contrastive pretext tasks to learn robust normal representations, and anomalies are detected as samples lying outside the distribution of contrastively trained features. These methods enhance the semantic structure of embeddings without relying on manual labels or artificial lesions.

Moreover, Some methods integrate anatomical priors such as brain symmetry. Since the healthy brain exhibits bilateral symmetry, anomalies often break this regularity. Symmetry-aware approaches, such as that of Ma *et al.* (2024), encode inter-hemispheric consistency using attention mechanisms to highlight deviations across hemispheres, enhancing sensitivity to focal anomalies.

Each approach offers a distinct view of what constitutes a “normal” embedding and how deviations from it are interpreted. These foundational differences drive their effectiveness and suitability across different types of brain abnormalities.

### 1.3.2 Synthetic-Based UAD Methods

Synthetic-based methods approach unsupervised anomaly detection by introducing artificial abnormalities (lesions) into healthy brain MR images, thereby converting the task into a supervised segmentation problem. These pseudo-anomalies serve as training targets, enabling the model to learn how to detect abnormal patterns in a fully supervised manner, despite never seeing real pathological labels.

One representative approach is *AutoSeg* (Meissen, Kaissis & Rueckert, 2021) which generates synthetic anomalies by inserting foreign textures into healthy brain MRIs using randomly shaped polygonal masks. The masked regions are filled by interpolating image patches from different patients, simulating realistic variations in lesion shape and texture. A U-Net is then trained to segment these inserted anomalies and regress their interpolation strength. Tan *et al.* (2022) proposed a self-supervised method based on "foreign patch interpolation". Their framework creates synthetic anomalies by inserting a patch from one healthy image into another and linearly blending its intensity with the surrounding tissue. Then, a network is trained to predict the location of these interpolated regions at the pixel level. This simple approach forces the model to detect subtle inconsistencies in texture and intensity, simulating how a real lesion might stand out from normal tissue.

To improve robustness, Baugh *et al.* (2023) proposed an ensemble approach that incorporates *multiple synthetic anomaly generation strategies*. Their method introduces a diverse set of defects, including Poisson blending, geometric warping, and intensity perturbations, during training. A single segmentation model is trained across all synthetic tasks using a unified anomaly labeling scheme. By exposing the model to a wider range of pseudo-anomalies, this approach improves generalization and enhances the model’s ability to detect real lesions.

Marimont *et al.* (2024) introduced *DISYRE* (Diffusion-Inspired Synthetic Restoration), a novel approach that integrates synthetic anomaly generation with diffusion-inspired restoration. Instead of using Gaussian noise for corruption, DISYRE applies gradual synthetic anomaly corruptions to healthy images. A restoration model is then trained to revert these corrupted images back to their original healthy state. The learned restoration function serves as an implicit anomaly detector, highlighting regions that deviate from the normative distribution. Building upon DISYRE, the same authors proposed "Ensembled Cold-Diffusion Restorations" (Naval Marimont *et al.*, 2024) which employs a cold-diffusion pipeline and introduces a novel synthetic anomaly generation procedure called DAG. By training the model to restore these synthetically corrupted images to their original form and ensembling restorations conditioned on varying degrees of abnormality, the approach enhances robustness and accuracy.

Synthetic-based methods offer the advantage of providing strong, pixel-level supervision during training, enabling precise, sharp anomaly localization without relying on real pathological annotations. However, the effectiveness of these methods is heavily contingent upon the realism and diversity of the synthetic anomalies. If the simulated defects are overly simplistic or visually distinct from true pathology, models may overfit to artificial cues, thereby impairing their ability to generalize to real-world anomalies. Consequently, while synthetic-based approaches can enhance localization accuracy, they inherently contrast with the principal advantage of unsupervised anomaly detection, namely, the capacity to identify a wide range of unforeseen anomalies without prior knowledge or explicit labels.

### 1.3.3 Reconstruction-Based UAD Methods

Reconstruction-based methods form a foundational class of approaches in unsupervised anomaly detection, particularly in brain MRI. The central premise is straightforward: by learning to reconstruct healthy anatomical structures, a model trained solely on normal images should struggle to accurately reproduce pathological regions during inference. Consequently, discrepancies between the input and reconstructed image can be interpreted as markers of abnormality. This

approach aligns with the intuition that anomalies cannot be faithfully reconstructed and will manifest as high reconstruction errors.

These methods are attractive due to their simplicity, data efficiency, and interpretability. Reconstruction-based strategies are widely used and served as the foundation for many state-of-the-art techniques, including those based on Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and, more recently, diffusion models. The following subsections review reconstruction-based methods categorized by their generative modeling framework, VAEs, GANs, and diffusion-based architectures, highlighting their methodological differences. Proposed UAD methods in this thesis, fall into the category. In particular, MAD-AD and DeCo-Diff are reconstruction-based methods that leverage diffusion models for anomaly detection, while REFLECT builds upon this foundation by replacing the diffusion process with a rectified flow formulation, offering improvements in both reconstruction fidelity and sampling efficiency.

#### 1.3.3.1 Reconstruction-Based UAD Methods Using VAEs

Variational Autoencoders (VAEs) provide a probabilistic generative framework that learns to model the distribution of healthy brain anatomy by reconstructing images from a low-dimensional latent space. In the context of anomaly detection, a VAE trained exclusively on healthy data is expected to reconstruct normal tissue faithfully, while failing to accurately reproduce pathological regions, which are then highlighted via reconstruction error.

A prominent early work in this direction is the context-encoding VAE (ceVAE) introduced by Zimmerer et al (Zimmerer *et al.*, 2018). Their model reconstructs masked regions of brain MRIs to exploit contextual dependencies, under the assumption that lesions disrupt local coherence. Anomalies are localized by measuring inconsistencies between predicted and original content in the masked areas. Uzunova *et al.* (2019) proposed a conditional VAE framework that separates anatomical variation from potential abnormalities by conditioning the latent representation on healthy priors. This conditional structure helps the model focus on deviations from normative brain structure, making it more robust to anatomical variability. Also, Baur *et al.* (2020c)

proposed to use a multi-resolution reconstruction strategy, using scale-space decomposition to improve the detection of anomalies across different spatial scales.

To further increase the accuracy of restoration of normal content in lesioned areas, Chen et al (Chen *et al.*, 2020b) introduced a restorative VAE trained to inpaint potentially anomalous regions with healthy reconstructions guided by normative priors. Unlike traditional VAEs, their model employs an adversarial loss in latent space to encourage separation between pathological and healthy features. A complementary strategy is proposed by Sato *et al.* (2019), who incorporate uncertainty estimation into the VAE framework. Their model produces multiple stochastic reconstructions using dropout-based sampling and identifies anomalies based on high reconstruction variance. This approach not only detects abnormal regions but also provides a pixel-wise confidence map, offering an additional interpretability layer.

To enhance reconstruction fidelity and robustness, Akrami et al (Akrami *et al.*, 2020) integrated transfer learning into the VAE pipeline. Their formulation is designed to reduce sensitivity to outliers or mislabeled healthy data during training, making it well-suited for real-world datasets where subtle anomalies may go unnoticed. Silva-Rodríguez, Naranjo & Dolz (2022) introduced a constrained UAD method by integrating inequality constraints into the VAE training process to homogenize the activations produced in normal samples, thereby enhancing the model's sensitivity to anomalies. By employing an extension of the log-barrier method and maximizing the Shannon entropy of attention maps, their formulation reduces the reliance on hyperparameter tuning and does not require access to anomalous images during training.

Despite their conceptual elegance and probabilistic formulation, VAE-based methods often struggle with practical limitations. A well-known issue is their tendency to produce overly smooth or blurry reconstructions, which can obscure subtle anomalies and reduce localization accuracy. Additionally, the reconstruction error may be spread across both normal and abnormal regions, especially when the latent space is not well-structured, leading to false positives or poor contrast between healthy and pathological tissue. These challenges limit their effectiveness in high-precision clinical applications.

#### 1.3.3.2 Reconstruction-Based UAD Methods using GANs

Generative Adversarial Networks (GANs) have been employed in unsupervised anomaly detection to improve upon the blurry reconstructions often associated with VAEs. By introducing an adversarial loss that encourages outputs to appear more realistic, GAN-based models enhance the visual sharpness and fidelity of reconstructed images. This increased fidelity can, in turn, lead to better delineation of anomalous regions when comparing the input and reconstructed outputs.

Although not originally designed for brain MRI, AnoGAN (Schlegl *et al.*, 2017) and its successor f-AnoGAN (Schlegl *et al.*, 2019b) were among the first GAN-based methods proposed for unsupervised anomaly detection in medical imaging and were later widely adopted for brain MRI applications. AnoGAN learns to generate healthy images by mapping inputs to a latent space trained exclusively on normal data. At inference time, anomaly detection is performed via an iterative optimization process that searches for the latent code yielding the closest reconstruction of a test image. The discrepancy between the input and its reconstruction serves as the anomaly signal. While effective, this iterative process is computationally expensive. To address this, f-AnoGAN introduced an encoder network that enables direct inference of latent representations, significantly accelerating the detection pipeline while preserving reconstruction quality.

In the context of brain MRI, Baur *et al.* (2021b) demonstrated that adversarial training can substantially enhance segmentation quality of VAE-based models by producing sharper reconstructions of healthy anatomy. To mitigate a common issue in GAN-based anomaly detection, namely, the tendency of models to unintentionally reconstruct anomalous regions (known as *anomaly leakage*), Chen & Konukoglu (2018) proposed adversarial constraints on both the image and feature level to explicitly discourage the generator from reproducing lesions. Similarly, Baur *et al.* (2020a) addressed the risk of *steganographic leakage* in CycleGAN-based methods, where lesions could be subtly encoded into the translated output without explicit reconstruction, thereby masking anomalies. Their proposed solution introduces anomaly suppression constraints to prevent such hidden leakage during domain translation.

A related line of work leverages inpainting as a means of localized anomaly detection. Nguyen *et al.* (2021) and Han *et al.* (2021) both proposed GAN-based models that reconstruct masked regions of brain MRIs conditioned on the surrounding healthy context. The idea is that a model trained solely on healthy data will inpaint masked lesions with plausible normal tissue, leading to large reconstruction errors at abnormal sites. These inpainting-based strategies support more spatially localized anomaly detection, as the mask confines the reconstruction task to specific regions.

While GAN-based models offer sharper reconstructions than VAEs and can produce visually compelling results, they also introduce challenges related to training instability, sensitivity to hyperparameters, and mode collapse. Moreover, adversarial training does not inherently guarantee that anomalies will be faithfully suppressed, making the careful design of architectural and loss components crucial for reliable detection.

#### 1.3.3.3 Reconstruction-Based UAD Methods using Diffusion Models

Diffusion models have recently emerged as a powerful alternative for high-fidelity image reconstruction, offering significant advantages over traditional VAE and GAN-based approaches. These models, particularly denoising diffusion probabilistic models (DDPMs) (Ho *et al.*, 2020b), learn to generate clean images by gradually denoising inputs from random noise. In the context of anomaly detection, diffusion models are trained exclusively on healthy images; at inference time, they aim to reconstruct a healthy version of potentially anomalous inputs. Deviations between the input and reconstructed image are then interpreted as potential anomalies.

Wolleb *et al.* (2022) introduced one of the earliest applications of diffusion models to medical anomaly detection. Their method trains a DDPM on healthy brain MRI and chest X-ray images, then uses a classifier-guided denoising process to steer the output toward normative anatomy. Anomaly maps are produced by comparing the original input with its denoised reconstruction, revealing abnormalities such as tumors or effusions. This approach demonstrated fine-grained localization and competitive results across modalities.

To improve efficiency and robustness, AnoDDPM (Wyatt *et al.*, 2022b) which introduced a partial noising scheme and replaces standard Gaussian corruption with Simplex noise. This noise type better destroys large pathological structures while preserving normal anatomical context, leading to more accurate anomaly suppression and reconstruction. AnoDDPM showed strong performance on brain MRI tumor segmentation tasks, outperforming previous GAN-based methods in both Dice score and inference stability.

pDDPM (Behrendt *et al.*, 2024b) advanced this line of work by proposing a patch-wise reconstruction approach. Rather than denoising the entire image at once, their model processes local patches independently, conditioned on surrounding uncorrupted context. This patch-wise inpainting allows for sharper and more anatomically consistent outputs, reducing the risk of anomaly leakage, a common issue in global reconstructions. Behrendt *et al.* (2024c) proposed leveraging the Mahalanobis distance to enhance unsupervised brain MRI anomaly detection. Their approach constructs multiple reconstructions using probabilistic diffusion models and analyzes the resulting distribution with the Mahalanobis distance to identify anomalies as outliers.

mDDPM (Iqbal *et al.*, 2023) introduced a masked diffusion model for unsupervised anomaly detection in medical images. By incorporating masking-based regularization, specifically Masked Image Modeling (MIM) and Masked Frequency Modeling (MFM), their self-supervised approach enables the model to learn visual representations from unlabeled data. This method demonstrated superior performance compared to existing fully/weakly supervised baselines on datasets containing tumors and multiple sclerosis lesions.

IterMask2 (Liang *et al.*, 2024a) which is a diffusion-inspired iterative anomaly segmentation framework, applies spatial and frequency-based masking to input images, reconstructs the masked regions using a model trained on healthy data, and iteratively updates the mask based on reconstruction errors. This iterative refinement improves segmentation accuracy by progressively isolating anomalous regions.

Further refinements have focused on guiding the denoising process to prevent hallucinations and preserve healthy context. THOR (Bercea *et al.*, 2024a) integrates temporal guidance maps into the diffusion process, enabling the model to selectively replace only abnormal regions. This strategy enhances anatomical fidelity and reduces false positives, particularly in datasets involving stroke lesions.

While diffusion models produce sharper and more reliable reconstructions than VAEs, and are more stable than adversarially trained GANs, they come with certain limitations. Inference time is a key bottleneck, as denoising involves a lengthy iterative process. Though recent work has introduced faster sampling strategies and adversarial guidance (Yu, Oh & Yang, 2023), real-time applicability remains a challenge. Additionally, overfitting to the training distribution of healthy images may cause false positives when encountering unseen but benign variations.

Despite these challenges, diffusion-based reconstruction methods currently represent the state-of-the-art in unsupervised medical anomaly detection. Their ability to produce high-quality, context-aware reconstructions with minimal supervision positions them as a promising direction for future research and clinical deployment.

In summary, existing unsupervised anomaly detection methods in brain MRI have demonstrated the potential of embedding-, synthetic-, and reconstruction-based frameworks for identifying pathological deviations without labeled data. However, many approaches suffer from low detection performance, limited generalization to rare or subtle anomalies, reconstruction-induced hallucinations, false identification of normal regions due to inaccurate reconstructions. To address these limitations, this thesis introduces three complementary approaches based on generative models: Chapter 3 presents MAD-AD, which leverages masked diffusion models to improve anomaly detection performance using selective correction; Chapter 4 proposes a reformulation of diffusion models to correct abnormalities as a form of *deviation from normality*, enabling more complex multi-class scenarios that can be applied to different modalities and organs; and Chapter 5 introduces REFLECT, which achieves strong performance in brain MRI unsupervised anomaly detection through an efficient one-step correction.

## 1.4 Neonatal Brain Age Estimation

A range of imaging-based methods have been proposed to estimate fetal or neonatal brain age (gestational or postmenstrual age) from MRI. Early approaches relied on volumetric and morphometric features. For example, Hüppi *et al.* (1998) showed that neonatal gray matter volume is highly correlated with postmenstrual age (PMA) at scan. Subsequent studies demonstrated that cortical folding metrics also track age: the fraction of cerebral surface occupied by sulci, sulcal depth, and curvature all increase with gestational age. Galdi *et al.* (2020) combined such features across modalities by constructing morphometric similarity networks from regional volumes and diffusion metrics. Their regression model predicted neonatal PMA at scan with a mean absolute error (MAE) of approximately 0.70 weeks and could also classify preterm versus term infants with 92% accuracy.

Diffusion MRI provides complementary information about white matter maturation and connectivity, which has also been leveraged for age prediction. Several works build structural connectomes from neonatal DTI and feed them into machine learning models. Zhao, Cai & Liu (2024) used both T2-weighted and diffusion MRI in a transformer-based model, achieving MAE near 0.5 weeks and revealing systematic delays in predicted brain age for preterm-born neonates. Sun *et al.* (2024) combined structural and functional connectomes to predict PMA and found that deviations from the normative brain age (brain age gap) were associated with preterm birth and neurodevelopmental risk.

Deep learning has become increasingly prominent in fetal and neonatal brain age estimation, often outperforming traditional models based on handcrafted features. In fetal MRI, convolutional neural networks (CNNs) with attention mechanisms have achieved high predictive accuracy. Shi *et al.* (2020) trained an attention-guided ResNet on fetal T2-weighted slices to estimate gestational age (15-39 weeks), achieving MAE $\approx$ 0.77 weeks and demonstrating the utility of uncertainty quantification as an anomaly marker. Shen *et al.* (2022) employed a multi-planar ResNet-50 with learned attention masks and achieved MAE $\approx$ 0.96 weeks on a cohort of 741 normal fetuses.

In neonatal MRI, both 3D CNNs and hybrid transformer models have been proposed. Chen *et al.* (2022) used relatively shallow 3D CNNs trained with ranking losses to regress PMA. Zhao *et al.* (2024) employed a dual-stream CNN-transformer architecture that integrates T2-weighted and diffusion images, achieving high accuracy and identifying maturational delays in preterm infants. Notably, many deep learning models produce interpretable attention or saliency maps; for instance, Shi *et al.* (2020) and Shen *et al.* (2022) visualized anatomical regions most influential in their models' predictions. These end-to-end models learn complex morphological patterns—such as cortical folding, ventricular volume, and tissue contrast—without relying on prior segmentation or anatomical atlases.

A number of studies have examined how prematurity impacts brain maturation and predicted brain age. Galdi *et al.* (2020) showed that the structural connectivity patterns predictive of PMA differ between term and preterm groups. Zhao *et al.* (2024) and Sun *et al.* (2024) explicitly found that preterm infants exhibit delayed predicted brain ages relative to chronological age, and that brain age gaps are significantly correlated with perinatal risk factors. These findings are consistent with prior literature demonstrating that preterm birth is associated with disrupted cortical development and altered structural network efficiency.

Across both traditional and learning-based approaches, several anatomical features have consistently emerged as correlates of brain age. Gray matter volume, cortical surface area, and gyrification index increase with age, while ventricular size typically decreases. White matter diffusion properties such as fractional anisotropy (FA) rise with maturation due to myelination and increased fiber coherence. Numerous studies, including those using feature importance or saliency analyses, have identified the frontal and occipital lobes, ventricles, deep nuclei, and major white matter tracts as key contributors to age prediction. In particular, morphometric and functional changes in these regions provide strong, quantifiable signatures of brain development, supporting their role as imaging biomarkers of maturation.

Chapter 6 proposes a learning-based neonatal age estimation method that offers both the high performance of recent deep learning approaches and the interpretability of traditional methods, while remaining effective with limited data.

# CHAPTER 2

# HARMONIZING FLOWS: LEVERAGING NORMALIZING FLOWS FOR UNSUPERVISED AND SOURCE-FREE MRI HARMONIZATION

Farzad Beizaee[1,2,3] , Gregory A. Lodygensky[3,4] , Chris L. Adamson[5] , Deanne K. Thompson[5,6,7] , Jeanie L.Y. Cheong[5,7,8,9] , Alicia J. Spittle[6,8,10] , Peter J. Anderson[5,6] , Christian Desrosiers[1,2] , José Dolz[1,2]

[1] LIVIA, ÉTS, Montreal, Quebec, Canada
[2] ILLS , McGill - ETS - Mila - CNRS - Université Paris-Saclay - CentraleSupelec, Canada
[3] CHU Sainte-Justine, University of Montreal, Montreal, Canada
[4] Canadian Neonatal Brain Platform, Montreal, Canada
[5] Murdoch Children's Research Institute, Parkville, Victoria, Australia
[6] School of Psychological Sciences, Monash University, Clayton, Victoria, Australia
[7] Department of Paediatrics, The University of Melbourne, Victoria, Australia
[8] The Royal Women's Hospital, Melbourne, Parkville, Victoria, Australia
[9] Department of Obstetrics and Gynaecology, The University of Melbourne, Victoria, Australia
[10] Department of Physiotherapy, The University of Melbourne, Victoria, Australia

Article published in the journal "Medical Image Analysis (MedIA)", February 2025

## 2.1 Introduction

Magnetic Resonance Imaging (MRI) serves as an indispensable tool in modern medical diagnostics, enabling clinicians to obtain detailed insights into anatomical structures and pathological conditions. However, the inherent variability in MRI data acquisition protocols across different imaging sites poses significant challenges in achieving consistent and reliable image interpretation. This variability can stem from differences in scanner hardware, imaging parameters, and patient populations (Takao *et al.*, 2011), leading to inconsistencies in image appearance and potentially confounding downstream analysis. For instance, MRIs acquired from two different scanners or with different sets of parameters and configurations will often have noticeable appearance differences, which can be considered as domain shift. Therefore, pooling multi-centric clinic trials to address specific questions does not necessarily enhance statistical power, as the introduced variance may partially stem from non-clinical sources.

On the other hand, despite the considerable progress observed in deep learning, these models still face challenges in coping with distributional shifts. The performance of deep neural networks in fundamental visual problems, such as classification, segmentation, and regression, largely degrades when they are applied to data acquired under varied conditions, consequently limiting their broad applicability. Specifically, models trained on data from a specific site often struggle to achieve similar performance when applied to images from other centers.

To mitigate this challenge, image harmonization tackles the problem of distributional drifts by mapping images from one domain to another, aiming at transferring contrast characteristics across diverse datasets. MRI harmonization ensures the comparability of MRI data collected from different scanners, facilitating accurate and consistent analysis for multi-center studies involving diverse imaging datasets. However, many existing harmonization approaches rely on assumptions that could impede their feasibility and scalability in real-world applications. For example, some methods involve acquiring imaging data of the same anatomical targets from multiple sites or locations. These methods, often referred to as using *traveling subjects*, aim to identify and quantify the transformations needed to harmonize the data across different acquisitions settings (Dewey *et al.*, 2019; Durrer *et al.*, 2023). Another family of approaches requires access to the source images during harmonization or knowing the target domain in advance (Pomponio *et al.*, 2020; Modanwal *et al.*, 2020; Liu *et al.*, 2021; Cackowski *et al.*, 2023), which might not be feasible in practical scenarios. It is worth mentioning that in this context, the source domain refers to the domain that serves as the reference for harmonization, representing the desired appearance or distribution that other domains (referred to as target domains) are aligned to after the harmonization process. Furthermore, several of these harmonization strategies require annotated data associated with the downstream task (Delisle *et al.*, 2021; Dinsdale *et al.*, 2021; Karani *et al.*, 2021). This poses an additional challenge to the harmonization, as acquiring labeled data can be resource-intensive and time-consuming, especially when dealing with large datasets and dense tasks such as segmentation. Finally, many harmonization techniques require knowledge of the target domains during the training phase, despite the common occurrence of

unknown target domains in real-world scenarios. Based on the limitations exposed above, we present the following contributions in this work:

- We alleviate the aforementioned constraints on MR harmonization and introduce a novel harmonization approach that is *unsupervised*, *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$) and can cope with *unknown-domains* ($\mathcal{UD}$) without necessitating retraining for every target distribution. In fact, our approach only requires MRIs from one modality of the source domain during training, as opposed to existing approaches.
- Specifically, we propose to leverage a modern family of generative models, known as normalizing flows, which have proven to be highly effective in modeling data distributions for generative purposes.
- Alongside the methodological novelty of the proposed method, our empirical findings illustrate that it yields significant improvements over existing harmonization techniques, while effectively mitigating their limitations. More importantly, the comprehensive experimental section on multiple tasks and datasets demonstrates that the proposed approach successfully generalizes across target tasks and population demographics.

A preliminary conference version of this work has been presented at IPMI 2023 (Beizaee *et al.*, 2023). This manuscript provides a substantial extension of the conference version, which includes *i)* an extended literature review on methods addressing the problem of distribution shift, and a comprehensive empirical validation of the proposed approach, including *ii)* additional recent approaches for harmonization, *iii)* extensive ablation studies to validate our choices, *iv)* assessing the performance of our approach in multiple tasks and population demographics, *v)* including additional adult datasets in the experiments, *vi)* employing additional evaluation metrics to assess the performance from a harmonization standpoint, *vii)* and complimentary plots and results to better understand the overall performance of the different studied methods.

## 2.2 Related work

**Image harmonization.** In the medical domain, various methods have been proposed to harmonize images, with a particular focus on MRI data. Traditional post-processing procedures

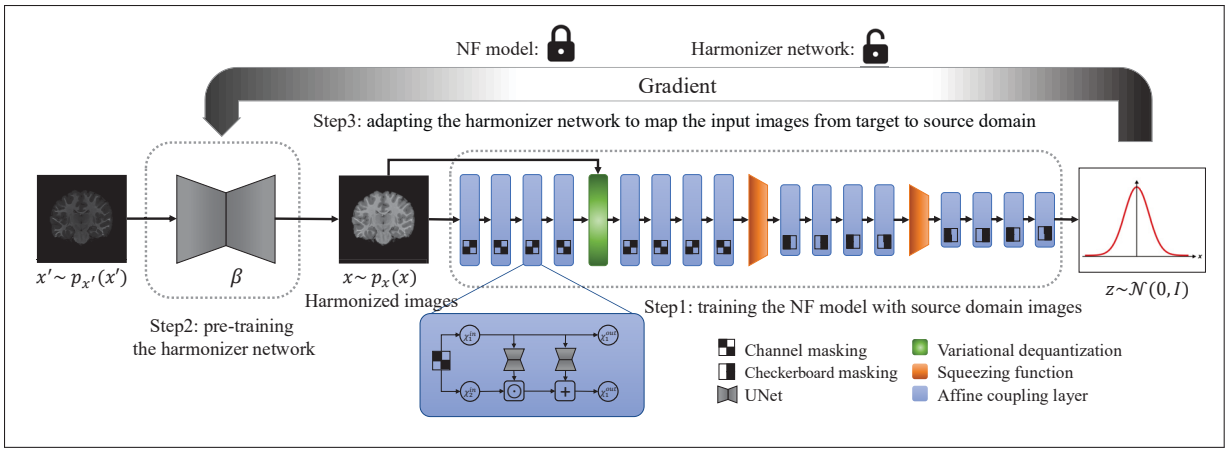



Figure 2.1 **Harmonizing Flows Pipeline.** Our method comprises three primary steps. Initially, normalizing flow (NF) is utilized to capture the distribution of the source domain. Subsequently, a harmonizer network undergoes pre-training to reconstruct the original images from augmented counterparts, facilitating initial harmonization. In the third stage (test time), the trained NF is leveraged to update the parameters of the harmonizer network, ensuring maximal alignment between the harmonized outputs and the learned NF distribution. Notably, steps 1 and 2 are independent of each other and can be executed interchangeably

like intensity histogram matching (Nyúl *et al.*, 2000; Shinohara *et al.*, 2014) help mitigate biases across scanners, but may also eliminate informative local intensity variations. Statistical harmonization approaches, on the other hand, can model both image intensity and dataset bias at the voxel level (Fortin *et al.*, 2016; Fortin *et al.*, 2017; Beer *et al.*, 2020). However, when the variations in data distribution are more complex and localized, they typically lead to sub-optimal harmonization outcomes. Additionally, these methods must often be adjusted each time images from new sites are provided, further compromising their performance in real-world applications. Modern strategies for image harmonization, using deep learning techniques, hold significant promise as an alternative solution for this problem (Dewey *et al.*, 2019; Zhu *et al.*, 2017; Liu *et al.*, 2021; Zuo *et al.*, 2021; Delisle *et al.*, 2021; Dinsdale *et al.*, 2021). Yet, these approaches often rely on unrealistic assumptions, which pose significant barriers to scalability when applied to extensive multi-site harmonization endeavors. First, some methods require images of the same target anatomy across different sites, known as *traveling subjects*, to identify intensity transformations among different sites (Dewey *et al.*, 2019; Durrer *et al.*, 2023). This means

that a given number of subjects being scanned at every site or scanner is employed for training, a condition rarely met in practice. The most widely used models for MR harmonization are GANs and autoecoders, which have shown promising results in reducing multi-site variation through image-to-image synthesis. GANs perform domain translation by learning domain invariant features. One issue of such models is that they are limited to mapping between two specific scanners for most studies. Also, target domains are required to be known at the training time (Zhu *et al.*, 2017; Liu *et al.*, 2021) which is a limiting factor for the scalability of the harmonization process. Additionally, each time a new domain is added, these approaches must be fine-tuned to accommodate the characteristics of this domain. Autoencoder-based methods (Torbati *et al.*, 2021; Dewey *et al.*, 2020; Zuo *et al.*, 2021, 2023; Wu *et al.*, 2023a; Cackowski *et al.*, 2023), on the other hand, aim to harmonize data in terms of disentangled representations. This group of harmonization methods attempted to extract scanner-related features for harmonization. Similar to GANs, data from multiple sources and target domains are required for training. CALAMITI (Zuo *et al.*, 2021), and HACA3 (Zuo *et al.*, 2023), which are two key methods of this category, require paired multi-modal MRIs for training, thereby restricting their applicability, particularly in single-modality scenarios. DLEST (Wu *et al.*, 2023a) and Imunity (Cackowski *et al.*, 2023) try to solve the problem of requiring multi-modal data for latent disentanglement. However, training these methods can be challenging due to the instability of their adversarial learning strategy. Lastly, task-dependent methods leverage labels associated with each image for a particular downstream task to optimize the harmonization for this specific problem (Delisle *et al.*, 2021; Dinsdale *et al.*, 2021). These methods often rely on task-specific features or assumptions, making them less effective when applied to new tasks or unseen data. Additionally, task-dependent harmonization approaches require large amounts of annotated data for training, which can be costly and time-consuming to acquire.

**Test-time Domain Adaptation.** Traditional solutions to the problem of distributional shift use labeled samples from a source domain and unlabeled ones from the target domain for adapting a source-trained model to perform well on the target. Several strategies for this task, known as unsupervised domain adaptation (UDA), work by explicitly aligning the feature distributions

of the source and target domains (Wang *et al.*, 2023b; Wu & Zhuang, 2020). Another popular approach consists in learning a domain-agnostic representation, for example using adversarial networks (Dou *et al.*, 2019; Kamnitsas *et al.*, 2017). Generative adversarial networks (GANs) can also be employed in style transfer methods to change the appearance of images from the target to the source while also preserving their semantic structures (Chen *et al.*, 2020a; Zhao *et al.*, 2021).

A major limitation of UDA methods is their need to have source examples during the adaptation phase, which may be impracticable in medical applications due to data sharing restrictions. Source-free domain adaptation (SFDA) approaches (Bateson *et al.*, 2022; Yang *et al.*, 2022; Stan & Rostami, 2024) relax this constraint and instead adapt the source-trained model using only unlabeled data from the target domain. However, these approaches are usually task-dependent and require an explicit adaptation process for each new target domain, hence are not compatible with the harmonization setting investigated in this work.

Another strategy for addressing distribution shift, more closely related to our method, is test-time adaptation (TTA) (Boudiaf *et al.*, 2022; Mummadi *et al.*, 2022; Liang *et al.*, 2020; Wang *et al.*, 2021a; Niu *et al.*, 2022a,b). Unlike SFDA, which performs adaptation in an offline step, this strategy adapts a pre-trained deep neural network to domain shifts encountered *during inference* on test samples. One of the earliest TTA approaches, called TENT (Wang *et al.*, 2021a), updates the normalization layers of the network by minimizing the Shannon entropy of predictions for test samples. In (Mummadi *et al.*, 2022), entropy minimization has been changed by optimizing a log-likelihood ratio and considering the normalization statistics of the test batch. EATA (Niu *et al.*, 2022a) instead introduces an active sample selection criterion to identify reliable and non-redundant samples. The model is then updated based on these samples to minimize entropy loss for test-time adaptation. In (Niu *et al.*, 2022b), authors propose a sharpness-aware and reliable entropy minimization method called SAR, further stabilizing the TTA process. Moreover, SHOT (Liang *et al.*, 2020) adapts the entire feature extractor with a mutual information loss, while using pseudo-labels to provide additional test-time guidance. Instead of updating the network parameters, LAME (Boudiaf *et al.*, 2022) uses Laplacian regularization to do a post-hoc

adaptation of the softmax predictions. Recent works have explored the potential of TTA for cross-site/modality segmentation of medical images. Authors of (Hu *et al.*, 2021) propose a TTA method for segmentation using a regional nuclear-norm loss to improve the discriminability and diversity of predictions and a contour regularization term to enforce segmentation consistency between nearby pixels. Contrary to this paradigm, where typically the target-task network (e.g., classification or segmentation) is adapted at inference based on surrogate losses on the network predictions, our work focuses on modifying the image appearance instead, which offers a more general solution, which is agnostic to the task at hand.

Furthermore, while (Karani *et al.*, 2021) uses the reconstruction error of an auto-encoder applied on segmentation outputs to normalize input images, it requires segmentation masks for the adaptation, which makes of this strategy task and annotation dependent.

**Normalizing flows.** Popular methods for generative tasks include generative adversarial networks (GANs) (Goodfellow *et al.*, 2014) and Variational Auto-encoders (Kingma & Welling, 2014). Despite their popularity and wide acceptance, these methods present several important limitations, including mode (Salimans *et al.*, 2016) and posterior collapse (Lucas *et al.*, 2019), training instability (Salimans *et al.*, 2016), and the incapability of providing an exact evaluation of the probability density of new data points. Recently, normalizing flows (NF) have emerged as a popular approach for constructing probabilistic and generative models due to their ability to model complex distributions (Dinh *et al.*, 2017). Normalizing flows involves mapping a complex distribution, often unknown or poorly characterized, to a simpler distribution, typically the standard normal distribution. This is accomplished through a series of invertible and differentiable transformations. These transformations allow for efficient density estimation, sampling, and generative modeling. One of the key advantages of normalizing flows is their ability to capture intricate dependencies within data while providing tractable likelihood estimation, enabling a wide range of applications across domains. While the majority of current literature has utilized NFs for generative purposes (e.g., image generation (Ho *et al.*, 2019; Kingma & Dhariwal, 2018), noise modeling (Abdelhamed, Brubaker & Brown, 2019), graph modeling (Zang & Wang, 2020)) and anomaly detection (Gudovskiy *et al.*, 2022; Kirichenko,

Izmailov & Wilson, 2020), recent findings also indicate their effectiveness in aligning a given set of source domains (Grover *et al.*, 2020; Usman *et al.*, 2020; Osowiechi *et al.*, 2023). Closely related to our problem, and up to the best of our knowledge, only a few attempts have investigated using normalizing flows to harmonize MR images. In particular, (Wang *et al.*, 2021b) presented a strategy that harmonizes pre-extracted features, i.e., brain ROI volume measures, and not image harmonization as in this work. In addition, extracting these ROIs requires pixel-wise labels, making of this approach task-dependent, contrary to our method which is task-agnostic. Furthermore, (Jeong *et al.*, 2023) is a concurrent approach that appeared after the conference version of this work was published. As discussed in their work, and shown empirically in our evaluation, although this method also leverages normalizing flows to harmonize images, it fails in the presence of large domain drifts between source and target sites.

## 2.3 Methodology

We first define the problem addressed in this study. Consider $\mathcal{X}_{\mathcal{S}} = \{\mathbf{x}_n\}_{n=1}^{N}$ as a collection of unlabeled images from the source domain $\mathcal{S}$, where each image $i$ is represented by $\mathbf{x}_i \in \mathbb{R}^{|\Omega|}$, with $\Omega$ indicating its spatial domain (i.e., $W \times H$). Likewise, let $\mathcal{X}_{\mathcal{T}} = \{\mathbf{x}_n\}_{n=1}^{M}$ be the set of unlabeled images within a target domain $\mathcal{T}$ [1]. The objective of unsupervised data harmonization is to discover a mapping function $f_\theta : \mathcal{T} \rightarrow \mathcal{S}$ without relying on labeled images or paired data from either domain.

We introduce a solution based on normalizing flows to address this problem. The proposed framework, which comprises three separate steps, is illustrated in Figure 2.1. In Step 1, we first utilize Normalizing Flows, renowned for their ability to accurately learn data likelihoods, to capture the source domain's distribution. In Step 2, we then employ an auto-encoder as a harmonizer network, pre-training it by reconstructing original images from the augmented source domain counterparts. Finally, during testing, we update the parameters of the harmonizer network using images from the unseen target domain, ensuring that the harmonized outputs

[1] For simplicity, we assume a single domain exists here. However, our formulation can be readily extended to accommodate $T$ distinct domains.

align with the learned distribution using NF model. The following sections present each of these steps in greater detail.

### 2.3.1 Learning the source domain distribution

We have leveraged normalizing flows (Dinh *et al.*, 2017) to model the source domain distribution. NFs are a modern family of generative models capable of modeling complex probability density $p_x(\mathbf{x})$ (i.e., the source domain distribution) through applying a sequence of transformation functions, denoted as $g_\phi = g_1 \circ g_2 \circ \dots g_T$, on known simple probability density $p_u(\mathbf{u})$ such as standard normal distribution. Source image can be represented as $\mathbf{x} = g_\phi(\mathbf{u})$, where $\mathbf{u} \sim p_u(\mathbf{u})$ and $p_u(\mathbf{u})$ denotes the base distribution of the normalizing flow model. An essential condition for the transformation function $g_\phi$ is its requirement to be *invertible*, with both $g_\phi$ and $g_\phi^{-1}$ being *differentiable*. With these prerequisites met, the density of the original variable $\mathbf{x}$ is well-defined, allowing for the exact computation of its likelihood using the change of variables rule, expressed as:

$$\begin{aligned} \log p_{\mathrm{x}}(\mathbf{x}) &= \log p_{\mathrm{z}}\left(g_\phi^{-1}(\mathbf{x})\right) + \log\left|\det\left(\mathbf{J}_{g_\phi^{-1}}(\mathbf{x})\right)\right| \\ &= \log p_{\mathrm{z}}\left(g_\phi^{-1}(\mathbf{x})\right) + \sum_{t=1}^{T} \log\left|\det\left(\mathbf{J}_{g_t^{-1}}(\mathbf{u}_{t-1})\right)\right| \end{aligned} \tag{2.1}$$

The first component of the right side corresponds to the log-likelihood within the simple distribution and $\mathbf{J}_{g_t^{-1}}(\mathbf{u}_{t-1})$ indicates the Jacobian matrix corresponding to the transformation $g_t$. In order to train the Normalizing flow model and learn the source domain distribution, model parameters $\phi$ are learned by maximizing the likelihood of the transformed data under the simpler distribution. This is achieved by minimizing the negative log-likelihood in Eq. 2.1 which leads to the following loss function:

$$\mathcal{L}_{NF} = -\log p_{\mathrm{x}}(\mathbf{x}) \tag{2.2}$$

**Building the Normalizing Flow.** Constructing a bijective transformation function neural network for the Normalizing Flow (NF) model often involves the stacking of affine coupling layers, as highlighted by (Dinh *et al.*, 2017; Kingma & Dhariwal, 2018). This approach has been established as an efficient strategy. Coupling layers offer computational symmetry, meaning

they are equally rapid in both evaluation and inversion processes. This characteristic addresses usability concerns inherent in asymmetric flows, such as masked autoregressive flows, which makes the coupling layers a preferred choice. Their balanced computational efficiency enables smoother integration into various applications, contributing to their widespread adoption in NF architectures. Suppose $\mathbf{z} \in \mathbb{R}^D$ serves as the input to the coupling layer, which is partitioned disjointly into $(\mathbf{z}^A, \mathbf{z}^B) \in \mathbb{R}^d \times \mathbb{R}^{D-d}$. The partitioning can be done along spatial dimensions(e.g. checkerboard masking strategy) or channels (channel masking strategy). Then, the transformation function $g(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^D$ can be expressed as:

$$\mathbf{y}^A = \mathbf{z}^A, \quad \mathbf{y}^B = \mathbf{z}^B \odot \exp\left(s\left(\mathbf{z}^A\right)\right) + t\left(\mathbf{z}^A\right) \tag{2.3}$$

where $\mathbf{y}^A$ and $\mathbf{y}^B$ represent the transformed parts of the input data and $\odot$ is element-wise multiplication. This formulation divides the input into two parts ($\mathbf{z}^A$ and $\mathbf{z}^B$), and transforms only the latter part, leaving the former unchanged. This setting offers simplicity for calculating the Jacobian determinant, which makes it possible to use complex neural networks as shift $s(\cdot)$ and scale $t(\cdot)$ networks. Note that the transformation in Eq. 2.3 is invertible and therefore allows for efficient recovery of $\mathbf{z}^A$ and $\mathbf{z}^B$ from $\mathbf{y}^A$ and $\mathbf{y}^B$. The work in (Dinh *et al.*, 2017) presented coupling flows on simpler tasks and datasets which demanded less complex representations. However, our current task necessitates pixel-to-pixel mappings on more challenging data. Therefore, we replace the simple convolutional blocks in (Dinh *et al.*, 2017) with shallow U-shaped convolutional neural networks to find the scale and shift parameters of the affine transformation, as they capture broader contextual information and provide higher representational capacity. Moreover, given that NFs rely on the change of variables rule, which operates within continuous space, it is important to ensure that the input is continuous. Traditionally, dequantization involves adding uniform noise $u \in U[0, 1]$ to discrete values to convert them into continuous representations. However, it could lead to a hypercube representation with sharp borders. Such sharp borders pose a challenge for modeling with a flow, as it relies on smooth transformations. Recently, a variational framework was introduced (Ho *et al.*, 2019) to expand dequantization to more sophisticated distributions. This was achieved by substituting the uniform distribution

with a learnable distribution. This learnable distribution can be optimized alongside other parameters of the normalizing flow, allowing for end-to-end training and seamless integration into the density estimation process.

**Constraining the source-distribution learning.** Optimizing the objective in Eq. 2.2 solely with source images could potentially bias the model towards emphasizing characteristics of subjects, such as age and gender, rather than focusing on source-specific attributes like contrast and brightness. To address this concern, we propose a strategy aimed at facilitating the learning of the source domain distribution. For this purpose, in each iteration, we randomly select $N'$ images of the source dataset $\mathcal{X}_S$ and apply a series of augmentations $f_{aug}(\cdot)$ in such a way that the resulting image exhibits a dissimilarity in appearance compared to the original image, as measured by the mean squared distance, surpassing a predetermined threshold. These images can be served as out-of-distribution samples, to guide the normalizing flow model to learn source-specific characteristics. In particular, we employ different types of contrast augmentation, brightness changes, multiplicative transformations, and random monotonically increasing mapping functions to augment these images. Then, the overall learning objective of our model can be defined as follows.

$$\mathcal{L}_T = \underbrace{-\sum_{n=1}^{N-N'} \log p_{\mathrm{x}}(\mathbf{x}_n)}_{\text{Source distribution modeling}} \underbrace{-\sum_{n=1}^{N'} \min\left(\mathrm{c}, -\log p_{\mathrm{x}}(f_{aug}(\mathbf{x}_n))\right)}_{\text{Guiding term}}. \tag{2.4}$$

The first term is the learning objective in Eq. 2.2 over the source images, while the second one encourages the NF model to reduce the likelihood of the augmented images, which facilitates the learning of domain-specific characteristics rather than subject-related features. Furthermore, to prevent divergence of the negative log-likelihood for an augmented sample to infinity, we employ a constant margin in the second term denoted as c.

### 2.3.2 Achieving image harmonization

**Harmonizer network.** The goal of our harmonizer network $h_\theta(\cdot)$ is to perform image-to-image translation of MRIs from the target to the source domain in such a way that: $p_{\mathbf{x}}(\mathbf{x}) = p_{\mathbf{x}'}(h_\theta(\mathbf{x}'))$. It's important to highlight that the input and output of the harmonizer network must share the same spatial dimensions. Here $\theta$ denotes the set of learnable parameters within the harmonizer network, and $\mathbf{x}$ and $\mathbf{x}'$ represent the samples from the source domain and target domains, respectively. This implies that the harmonizer network aims to shift the distribution of the target images so that they align with the distribution of source images. Nonetheless, we want the proposed method to operate effectively on unseen domains, necessitating that the target domains remain unknown during training. To this end, at first, the harmonizer network was trained to restore the original MRIs of the source domain from its augmented version. As in the previous step, we used different types of contrast augmentation, brightness changes, multiplicative transformations, and random monotonically increasing mapping functions. In contrast to the first step, there are no restrictions on how much the image can be altered as long as the augmented image is logical and details are not eliminated. To train this model, we employed the sum of two commonly used standard reconstruction loss functions: SSIM (Structural Similarity Index Measure) loss (Wang *et al.*, 2004) and L1 loss (mean absolute error). SSIM loss is more appropriate when preserving structural similarity and perceptual quality is important, while L1 loss is suitable for tasks where exact pixel-wise accuracy is required, regardless of perceptual differences. The learning objective for the harmonizer network thus becomes:

$$\theta^{init} = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \Big( \sum_{n=1}^{N} \Big\| (\mathbf{x}_n - h_\theta \left( f_{aug}(\mathbf{x}_n) \right) \Big\| + SSIM(\mathbf{x}_n, h_\theta \left( f_{aug}(\mathbf{x}_n) \right)) \Big) \tag{2.5}$$

A simple UNet has been considered as the harmonizer network. Also, it's important to emphasize that the conducted augmentations may not perfectly represent potential unseen target domains.

So, directly using the trained parameters $\theta^{init}$ for image-to-image translation yields sub-optimal harmonization. Nevertheless, it provides a good starting point for the next phase.

**Adapting the harmonizer network leveraging the Normalizing Flow.** So far, we have obtained an initial harmonizer network that gives us the possibility of transforming image appearance across domains, while bearing in mind that the output is sub-optimal. Also, we have learned the exact distribution of the source domain using a normalizing flow model, where we ensured that it focuses on domain-specific characteristics. The final step involves refining the harmonizer network to effectively map images from the target domain onto the distribution of the source domain. For this purpose, firstly, we stack the trained NF model at the top of the pre-trained harmonizer network as the Figure 2.1. Note that as we aim to leverage the learned distribution by the NF model which is already trained with the source data, its parameters should remain frozen during the adaptation of the harmonizer. Then we try to optimize the parameters of the harmonizer network, such that harmonizer network outputs exhibit a high likelihood of aligning with the source domain distribution under the supervision of the trained NF model. The learning objective of the adaptation stage is to increase the likelihood of the harmonizer outputs for images from the target domain, based on the the density estimation provided by the NF model. This objective is encapsulated in a defined loss function, which can be expressed as follows:

$$\mathcal{L}_{Adap} = -\sum_{m=1}^{M} \log p_{\mathrm{x}}\big(g_{\phi}\big(h_{\theta}(\mathbf{x}_m)\big)\big) \tag{2.6}$$

As a stopping condition for updating the harmonizer, we assess two potential alternatives: One criterion involves assessing the Shannon entropy of the predictions generated by the target downstream task using the harmonized images (e.g., classification or segmentation), and stopping the adaptation when the entropy plateaus. Additionally, we take into account the bits per dimension (*BPD*), which is a scaled variation of the *negative log-likelihood* commonly utilized for assessing generative models:

$$BPD = -\log p_{\mathrm{x}}(\mathbf{x}) \cdot \Big(\log 2 \cdot \prod_{i} \Omega_i\Big)^{-1} \tag{2.7}$$

where $\Omega_1$, ..., $\Omega_T$, are the spatial dimensions of the input images. More concretely, we can stop updating the harmonizer network when the reached *BPD* value matches the BPD observed for the source images using the trained NF model. In practice, the source BPD value can be determined during the training time using a validation set.

Table 2.1 **Acquisition parameters across different sites.** Scanner details and phenotypic information for each site used in this study. *y*: years; *gw*: gestation weeks

| Sites | # used MRIs | Scanner | TR (ms) | TE (ms) | Flip angle | Voxel size ($mm^3$) | Age |
|---|---|---|---|---|---|---|---|
| CALTECH | 19 | Siemens | 1590 | 2.73 | 10 | $1.0 \times 1.0 \times 1.0$ | 17–56 y |
| KKI | 20 | Phillips | 8 | 3.70 | 8 | $1.0 \times 1.0 \times 1.0$ | 8–13 y |
| NYU | 20 | Siemens | 2530 | 3.25 | 7 | $1.3 \times 1.0 \times 1.3$ | 6–39 y |
| PITT | 20 | Siemens | 2100 | 3.93 | 7 | $1.1 \times 1.1 \times 1.1$ | 9–35 y |
| HBNSI | 77 | Siemens | 2730 | 1.64 | 7 | $1.0 \times 1.0 \times 1.0$ | 5–21 y |
| HBNRU | 57 | Siemens | 2500 | 3.15 | 8 | $0.8 \times 0.8 \times 0.8$ | 5–21 y |
| OASIS | 117 | Siemens | 9.7 | 4.0 | 10 | $1.0 \times 1.0 \times 1.25$ | 18–96 y |
| DHCP T1-w | 280 | Philips | 4795 | 8.7 | N/A | $0.8 \times 0.8 \times 0.8$ | 37–45 gw |
| DHCP T2-w | 333 | Philips | 12000 | 156 | N/A | $0.8 \times 0.8 \times 0.8$ | 37–45 gw |
| V2LP T1-w | 122 | Siemens | 2100 | 3.39 | 9 | $1.0 \times 1.0 \times 1.0$ | 37–45 gw |
| V2LP T2-w | 122 | Siemens | 8910 | 152 | 120 | $1.0 \times 1.0 \times 1.0$ | 37–45 gw |

## 2.4 Experiments

### 2.4.1 Experimental setting

First, we resort to the cross-site brain MRI segmentation task to evaluate the harmonization performance of different methods. We chose this task as it allows us to assess not only how effectively the proposed method aligns target domain images with the source domain, but also to evaluate whether the structural details are preserved well during the harmonization process. Furthermore, we investigate how well the proposed approach performs across different populations, encompassing both neonates and adults, and imaging modalities (i.e., T1-weighted and T2-weighted MRI). Last, to demonstrate the generalizability of our method, we explore its performance on the distinct task of neonatal brain gestational age estimation.

**Datasets**

*MRI harmonization.* It is important to recall that even though the empirical validation showcases the results across all the available sites, the proposed model only has access to a unique domain, i.e., the source, during the training steps and a unique target domain during the harmonization step. The details of the different datasets used for each task are provided below.

*Adult brain MRI segmentation.* In the context of adult brain MRI segmentation, we utilized data from a total of seven different sites[2]. Four of these sites are drawn from the Autism Brain Imaging Data Exchange (ABIDE) (Di Martino *et al.*, 2014) dataset, which includes: California Institute of Technology (CALTECH), Kennedy Krieger Institute (KKI), University of Pittsburgh School of Medicine (PITT), and NYU Langone Medical Center (NYU) sets. Out of the remaining sites, Staten Island (SI) and Rutgers University (RU) sites are sourced from the Healthy Brain Network (HBN) (Alexander *et al.*, 2017) dataset, which we refer to as HBNSI and HBNRI in this paper, along with data from the Open Access Series of Imaging Studies (OASIS) (Marcus *et al.*, 2007). We selected T1-weighted MRIs of a healthy control population from each site, which were skull-stripped, motion-corrected, and quantized to 256 intensity levels. For each site, 60% of the images are used as the training set, 15% as the validation set, and the remaining 25% for testing, which are exploited in a 2D manner using the coronal plane slices. Also, the dimensionality of each 2D brain MRI slice is $256 \times 256$. with resolutions mentioned in Table 2.1. Moreover, following other large-scale studies (Dolz, Desrosiers & Ayed, 2018), we used Freesurfer (Fischl, 2012) to obtain the segmentations and grouped them into 15 labels: background, cerebral GM, cerebral WM, cerebellum GM, cerebellum WM, CSF, ventricles, brainstem, thalamus, hippocampus, putamen, caudate, pallidum, amygdala and ventral DC.

*Neonatal brain MRI segmentation and age estimation.* We employed T1-weighted and T2-weighted MRIs from the developing Human Connectome Project (DHCP) (Makropoulos *et al.*, 2018), VIBES2 (Spittle *et al.*, 2014) and LaPrem (Cheong *et al.*, 2021) datasets. As VIBES2 and LaPrem datasets are acquired with the same imaging device and parameters, we have

[2] Please note that in the conference version of this work (Beizaee *et al.*, 2023), only four datasets were employed in the experiments, which explains the differences in the empirical results from both versions.

combined them and considered them as one site, which we refer to as V2LP hereafter in this paper. All MRIs were sourced from a healthy control population aged between 37 and 45 weeks of gestational age. For neonatal brain MRI segmentation, we utilized the coronal view 2D slices, while the axial view 2D slices were chosen for brain age estimation due to their richer information content. Also, similar to adults, the dimensionality of 2D slices is $256 \times 256$. The preprocessing and data splitting procedures for neonates were consistent with those used for adults, with the only difference being the inclusion of contours of 35 regional structures obtained from M-CRIB-S (Adamson *et al.*, 2020). Image acquisition parameters and device, number of used MRIs, and population age can be found in Table 2.1 for both adults and neonatal datasets. These values showcase how the selected datasets have distinct imaging devices and parameters.

**Harmonization baselines.** The proposed method is compared to a set of harmonization and image-to-image translation approaches. First, we apply either the segmentation or age regression network directly on non-harmonized images, which we refer to as a "Baseline", so we can assess the improvement gained using the harmonized images. Furthermore, our comparison also includes the following harmonization strategies: Histogram Matching (Nyúl *et al.*, 2000), Combat (Pomponio *et al.*, 2020), SSIMH (Guan *et al.*, 2022), two popular generative-based approaches, i.e., Cycle-GAN (Modanwal *et al.*, 2020) and Style-Transfer (Liu *et al.*, 2021), a source free latent-disentanglement harmonization (Imunity) (Cackowski *et al.*, 2023), and a recent method for harmonization based on normalizing flows (BlindHarmony) (Jeong *et al.*, 2023).

**Test time domain adaptation and generalization baselines.** In addition to the existing harmonization methods, the proposed approach is benchmarked against several test time domain adaptation and generalization methods. These methods include: aleatoric uncertainty estimation (AUE) (Wang *et al.*, 2019), which uses test time augmentation to adapt to the target domain; BigAug (Zhang *et al.*, 2020), which uses heavy augmentations on MRIs for generalization; and TENT (Wang *et al.*, 2021a) and SAR (Niu *et al.*, 2022b), which are test-time adaptation methods based on the segmentation's output confidence, i.e., entropy. The comparison with

Table 2.2 **Performance overview on the cross-site adult MRI segmentation task.** Segmentation performance, in terms of DSC and HD95 metrics, across different harmonization approaches. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold**

| Method | Strategy | $\mathcal{SF}$ | $\mathcal{TA}$ | $\mathcal{UD}$ | DSC (%) | HD95 (mm) |
|---|---|---|---|---|---|---|
| Baseline | – | – | – | – | 36.3 ±4.8 | 43.1 ±5.3 |
| Hist matching (Nyúl *et al.*, 2000) | Harmonization | ✓ | ✓ | ✓ | 63.7 ±4.8 | 12.3 ±2.7 |
| Combat (Pomponio *et al.*, 2020) | Harmonization | ✗ | ✓ | ✗ | 73.0 ±4.3 | 6.3 ±2.6 |
| Cycle-GAN (Modanwal *et al.*, 2020) | Harmonization | ✗ | ✓ | ✗ | 75.0 ±2.8 | 5.5 ±1.7 |
| Style-transfer (Liu *et al.*, 2021) | Harmonization | ✗ | ✓ | ✗ | 70.8 ±5.7 | 8.2 ±2.6 |
| SSIMH $_{\text{MLMI'22}}$ (Guan *et al.*, 2022) | Harmonization | ✗ | ✓ | ✓ | 59.4 ±5.1 | 12.3 ±2.9 |
| ImUnity $_{\text{MedIA'23}}$ (Cackowski *et al.*, 2023) | Harmonization | ✗ | ✓ | ✓ | 58.2 ±4.6 | 17.0 ±3.3 |
| BlindHarmony $_{\text{ICCV'23}}$ (Jeong *et al.*, 2023) | Harmonization | ✓ | ✓ | ✓ | 62.2 ±6.1 | 13.7 ±3.2 |
| AUE (Wang *et al.*, 2019) | Test Time Augmentation | ✓ | ✗ | ✓ | 35.4 ±3.6 | 24.5 ±3.6 |
| BigAug $_{\text{TMI'20}}$ (Zhang *et al.*, 2020) | Generalization | ✓ | ✗ | ✓ | 82.0 ±2.1 | 3.2 ±0.8 |
| TENT $_{\text{ICLR'21}}$ (Wang *et al.*, 2021a) | Test Time Adaptation | ✓ | ✗ | ✓ | 72.8 ±3.4 | 7.4 ±2.2 |
| SAR $_{\text{ICLR'23}}$ (Niu *et al.*, 2022b) | Test Time Adaptation | ✓ | ✗ | ✓ | 70.9 ±4.0 | 9.9 ±2.6 |
| **Harmonizing Flows** | Harmonization | ✓ | ✓ | ✓ | **82.9 ±2.3** | **3.1 ±1.0** |

these methods aims to shed light on the importance of MRI harmonization and to reveal whether MRI harmonization can be replaced by domain generalization or test-time adaptation methods.

**Evaluation metrics and protocol.** *Segmentation.* To evaluate the performance of the proposed MRI harmonization approach on adult cross-site brain MRI segmentation, we train a segmentation neural network $S_\Phi(\cdot)$ using the training set of the source domain. Then for each target domain, the segmentation performance is evaluated using the harmonized images, where the segmentation performance is measured with the Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95). We repeat these steps considering each site as a source domain and the remaining sites as the target domains in each iteration, and then we report the average of these metrics. The evaluation on neonatal brain MRI segmentation follows the same protocol as that of the adults, which is repeated for each of the two modalities considered, i.e., T1-weighted and T2-weighted.

*Neonatal brain age estimation.* The neonatal brain gestational age estimation task follows the same procedure, except that instead of the segmentation model, a deep regression network has

been trained. Following the nature of the task, the metrics used to assess its performance are the Mean Absolute Error (MAE) and the Mean Square error (MSE).

*Harmonization performance.* Last, following (Parida *et al.*, 2024), we resort to the Wasserstein distance (WD) (Kantorovich, 1960) between the normalized intensity histograms of harmonized images and those of the source domain. The WD calculates the smallest "cost" to change one distribution into the other, taking into account both the amount of change needed and how far the changes need to be shifted. Thus, the Wasserstein distance between two normalized histograms tells us how different they are by considering not only the differences in terms of change magnitudes (such as the KL divergence does), but also how far apart the differing parts are. In our experiments, and as stated earlier, we consider one single site as a source domain (while all the remaining domains/sites remain unknown). Furthermore, we report the average of the WD for all source-domain pairs.

**Implementation details.**

*Normalizing Flow (FL).* The NF model has been trained for 20,000 iterations with Adam optimizer and a batch-size of 32, an initial learning-rate of $1 \times 10^{-3}$, and using a weight-decay of 0.5 every 2000 iterations. We employ a U-shaped architecture within the coupling layers, comprising four different scales with a scaling factor of 2. Each layer consists of an activation function followed by a convolutional layer and a normalizing layer. The activation function used in each scale is a modified version of ELU, i.e., concat[ELU($x$), ELU($-x$)], which makes it easier for the NF model to map to normal distribution due to its symmetry properties. Additionally, there are 16, 32, 48, and 64 kernels in each scale, respectively. To construct the NF model, we initially employ four sequential coupling layers with checkerboard masking, aimed at capturing the noise distribution through variational dequantization. Then, this is followed by four identical coupling layers and a squeezing function as explained in (Dinh *et al.*, 2017) to decrease the spatial dimension. Then, we subsequently incorporate four more coupling layers employing a channel-masking strategy, and an additional layer of feature squeezing, followed by a final series of four coupling layers employing channel-masking. The overall architecture of the flow

model is shown in Figure 2.1. Also, the margin c used for guiding the NF model (Eq. 2.4) is empirically set to 1.2. For adapting the harmonizer network using images of the target domain during test time, the Normalizing Flow model is frozen, and the harmonizer model is updated slightly using Adam optimizer with a learning rate of $5 \times 10^{-7}$, and a batch-size of 32 until the epoch where the stopping criterion has been reached. All the models were implemented on PyTorch using two NVIDIA RTX A6000 GPU cards.

*Harmonizer.* The utilized UNet network as harmonizer consists of five different scales with a scaling factor of 2, where each scale includes a layer of the modified ELU activation function, i.e., concat[ELU($x$), ELU($-x$)], followed by two convolutional layers. The number of kernels of the convolutional layers for each scale is 16, 32, 48, 64, and 64 respectively.

*Segmentation and regression models.* The segmentation network employed in our empirical validation is nn-UNet (Isensee *et al.*, 2021) with batch-normalization, stride 2, and kernel size of 3, whereas ResNet18 (He *et al.*, 2016) serves as the age estimation network. Furthermore, all the segmentation, age estimation, and harmonizer networks are trained for 5000 iterations with Adam optimizer with an initial learning rate of $1 \times 10^{-3}$, a weight decay of 0.5 every 500 iterations, and a batch-size of 64. Last, the model at the best iteration, based on an independent validation set, is utilized.

It is worth mentioning that we did not use any type of contrast or intensity augmentation in the training of segmentation and regression networks, to better capture the effect of the harmonization methods and make them more sensitive to distribution shifts.

### 2.4.2 Results

In this section, we reported the empirical results of the experiments performed. In particular, we first resorted to the cross-site brain MRI segmentation results to evaluate the performance of different harmonization approaches, as the segmentation task is a good indicator of harmonization performance. Following the main results, we conducted a series of comprehensive ablation studies to empirically support our choices. Then, to show the task-agnostic nature of our

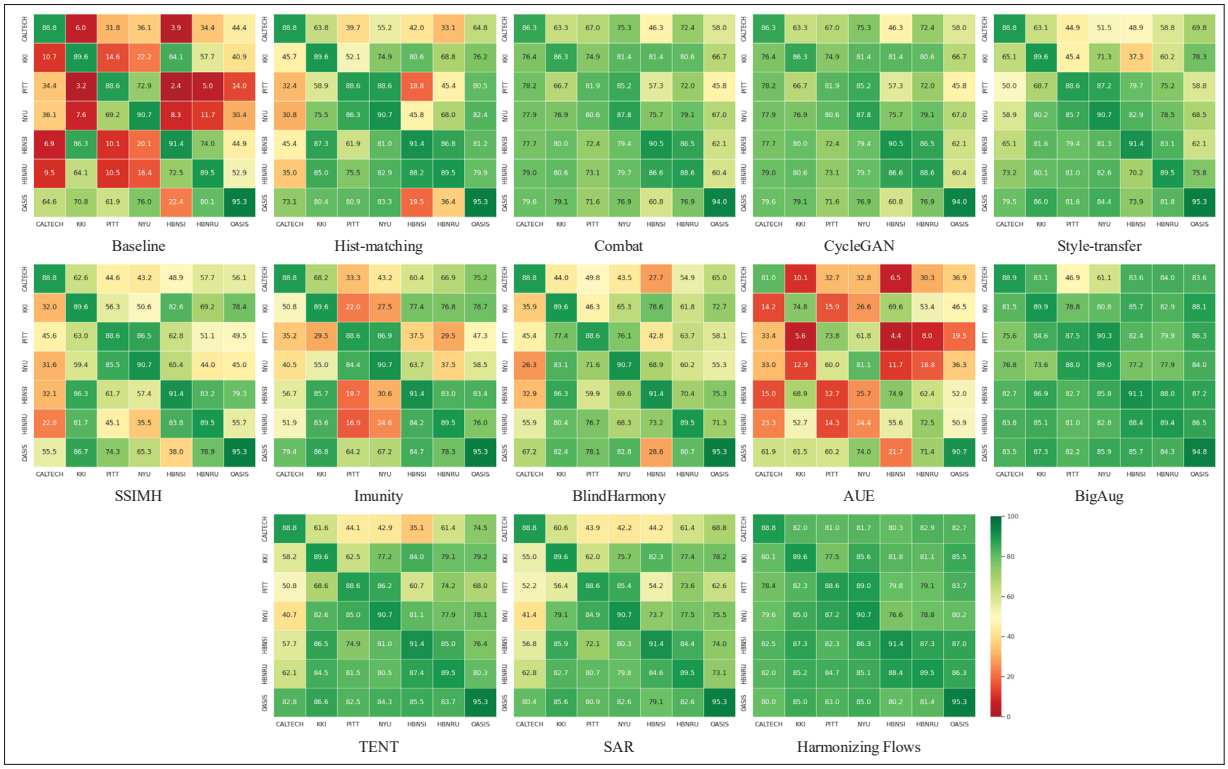



Figure 2.2 **Cross-site brain MRI segmentation matrix across the compared methods.** Each cell indicates the segmentation result (DSC %) when the source dataset (in the *rows*) is used to harmonize each target dataset (in *columns*)

method, we evaluated different harmonization strategies on the task of neonatal cross-site brain age estimation. Last, we took a closer look at the distance between the intensity histograms of the harmonized images with the source images. Surprisingly, closer intensity histograms between harmonized images with the source domain does not necessarily correlate with better harmonization performance or higher target task performance, such as segmentation or regression.

#### 2.4.2.1 Performance on the segmentation task

**Main results** To evaluate the proposed harmonization method, cross-site brain MRI segmentation performance has been obtained before and after the harmonization. As the segmentation networks are fixed, the segmentation results' improvement indicates the impact of the harmonization.

Segmentation results obtained with the images harmonized by different methods are reported in Table 2.2.

*Comparison to harmonization methods.* Before looking at the segmentation results, we would like to highlight that most existing methods rely on assumptions that could limit their scalability and practicality in real-life scenarios. First, some methods must access at least one image of the source domain during the adaptation, thereby not being completely *source-free* ($\mathcal{SF}$). Also, most harmonization techniques need to access the target domains during training, while ideally, the potential target domains should remain unknown, which we refer to as *unknown-domains* ($\mathcal{UD}$). We relax all these assumptions by proposing a method that is *source-free*, *task-agnostic*, and unaware of potential target domains during training. First, from Table 2.2, we can observe that the proposed approach improves the segmentation results by more than 45% in terms of DSC over the baseline, i.e., without harmonization. Furthermore, it consistently outperforms other harmonization approaches by a significant margin, in terms of both segmentation metrics. To statistically validate the performance of our method, we conducted a paired t-test between our method and each of the comparing harmonization methods. The test was performed using the results from different source datasets as the data samples for each comparison. Across all comparisons, the maximum p-value observed was p=0.012, which is below the commonly accepted significance threshold of 0.05, indicating that the observed improvements are consistently statistically significant. Particularly, the average gained improvement compared to CycleGAN, the next best-performing method, is larger than 8% in terms of DSC, and 2.4 *mm* smaller in terms of HD95. Considering that CycleGAN requires the source domain, as well as all the target domains to adapt, the differences in performance are even more important. Furthermore, if we compare it to approaches that offer the same benefits, i.e., ($\mathcal{SF}$), ($\mathcal{TA}$) and ($\mathcal{UD}$), such as Hist matching, and BlindHarmony these differences increase up to 20%.

*Comparison to test-time domain adaptation and generalization approaches.* To better evaluate the method, we have compared it against common strategies to tackle distributional shift, including: domain generalization (Zhang *et al.*, 2020), Test time Augmentation (Wang *et al.*, 2019), and Test time adaptation (Wang *et al.*, 2021a; Niu *et al.*, 2022b). This comparison

Table 2.3 **Performance overview on the cross-site neonatal MRI segmentation task.** Segmentation performance, in terms of DSC and HD95 metrics, across different harmonization approaches. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold**

| Method | $\mathcal{SF}$ | $\mathcal{TA}$ | $\mathcal{UD}$ | T1-w | | T2-w | |
|---|---|---|---|---|---|---|---|
| | | | | DSC (%) | HD95 (mm) | DSC (%) | HD95 (mm) |
| Baseline | – | – | – | 34.7 ±6.5 | 38.4 ±15.8 | 44.2 ±17.1 | 30.8 ±17.0 |
| Hist matching (Nyúl *et al.*, 2000) | ✓ | ✓ | ✓ | 42.9 ±8.0 | 32.9 ±17.6 | 84.3 ±4.1 | 4.7 ±2.3 |
| Combat (Pomponio *et al.*, 2020) | ✗ | ✓ | ✗ | 83.0 ±4.5 | 2.7 ±1.4 | 88.0 ±1.5 | 2.0 ±1.1 |
| Cycle-GAN (Modanwal *et al.*, 2020) | ✗ | ✓ | ✗ | 77.2 ±6.5 | 4.8 ±3.1 | 88.3 ±1.5 | 1.8 ±0.9 |
| Style-transfer (Liu *et al.*, 2021) | ✗ | ✓ | ✗ | 56.7 ±9.0 | 19.8 ±9.9 | 88.6 ±1.4 | 1.5 ±0.2 |
| SSIMH $_{\text{MLMI'22}}$ (Guan *et al.*, 2022) | ✗ | ✓ | ✓ | 38.7 ±9.7 | 30.2 ±13.3 | 61.9 ±8.5 | 17.9 ±6.6 |
| Imunity$_{\text{MedIA'23}}$ (Cackowski *et al.*, 2023) | ✗ | ✓ | ✓ | 41.8 ±7.9 | 23.8 ±9.5 | 87.1 ±3.0 | 2.4 ±1.8 |
| BlindHarmony $_{\text{ICCV'23}}$ (Jeong *et al.*, 2023) | ✓ | ✓ | ✓ | 56.1 ±9.2 | 28.4 ±12.6 | 88.0 ±1.8 | 2.9 ±1.6 |
| AUE (Wang *et al.*, 2019) | ✓ | ✗ | ✓ | 38.0 ±6.1 | 25.5 ±7.0 | 68.7 ±5.2 | 15.5 ±1.9 |
| BigAug $_{\text{TMI'20}}$ (Zhang *et al.*, 2020) | ✓ | ✗ | ✓ | 84.1 ±3.8 | 2.2 ±53.9 | **90.0 ±1.3** | **1.4 ±0.3** |
| TENT $_{\text{ICLR'21}}$ (Wang *et al.*, 2021a) | ✓ | ✗ | ✓ | 76.3 ±3.8 | 4.7 ±2.2 | 88.1 ±1.6 | 2.4 ±1.6 |
| SAR $_{\text{ICLR'23}}$ (Niu *et al.*, 2022b) | ✓ | ✗ | ✓ | 72.3 ±5.1 | 5.1 ±2.2 | 88.9 ±1.7 | 1.7 ±0.6 |
| **Harmonizing Flows** | ✓ | ✓ | ✓ | **84.4 ±2.5** | **2.1 ±1.0** | 89.6 ±1.4 | **1.4** ±0.4 |

highlights the importance of image harmonization and examines whether harmonization can be replaced by any of these strategies. These results, which are depicted at the bottom of Table 2.2, demonstrate that the proposed harmonization method also outperforms well-known test-time domain adaptation and generalization strategies in the task of cross-site brain segmentation. Furthermore, individual cross-site brain MRI segmentation results are depicted in Fig 2.2, for a better interpretability of the per-site results. In every matrix, the diagonal elements represent the segmentation of intra-site brain MRI and establish the upper bound of segmentation results when test images originate from the source domain. Then, the elements outside the diagonal indicate the segmentation result when a given dataset (indicated in the *rows*) is used to harmonize the target datasets (in *columns*). As the non-diagonal elements approach the diagonal, it can be interpreted as an enhanced capacity of the method to handle distributional shifts. Based on these results, we can state that our approach proved the most effective in this aspect, and consistently across all source-domain datasets.

**Validation on a different population and image modality**

To investigate the scalability of our harmonization method, we evaluate it using different population demographics and image modalities. Particularly, we resort to the neonatal brain MRI segmentation task, which differs from the more traditional adult brain MRI segmentation. Furthermore, this evaluation has been repeated for both T1-weighted and T2-weighted MRIs to explore whether the proposed approach can yield satisfactory performance for both modalities. According to the results from this experiment, which are reported in Table 2.3, our method demonstrates comparable performance for both modalities in neonatal brain MRI segmentation. More concretely, the proposed harmonization strategy obtains the best results in T1-weighted, whereas it ranks first among compared harmonization methods and second by a small margin among all compared methods for both metrics in T2-weighted images. These values underscore the effectiveness of our proposed approach across diverse populations and modalities.

**Ablation studies**

**I-Impact of normalizing flows.** This section assesses the impact of each component of the proposed method, which is achieved by comparing the cross-site brain MRI segmentation results obtained: *i)* when images are not harmonized, *ii)* when harmonized just with the proposed pre-trained harmonizer network $\theta^{init}$, or *iii)* harmonized with the proposed method. The results of this ablation study, which are depicted in Fig. 2.3, empirically support that the proposed NF-based models can serve as an effective mechanism to adapt the harmonizer network. First, the proposed approach to pre-train the harmonizer network yields significant enhancements compared to non-harmonized images (nearly 44% of DSC on average) while, despite its simplicity, there is no need to know the target domain in advance. Furthermore, adapting the harmonizer network using the proposed NF model further improves the cross-site MRI segmentation results, with nearly 2.0% of DSC on average, illustrating the effectiveness of the proposed harmonization strategy.

**II-Adaptation stopping criterion.** Now we explore the crucial issue of determining the appropriate iteration to stop the adaptation, specially as updating the harmonizer network is

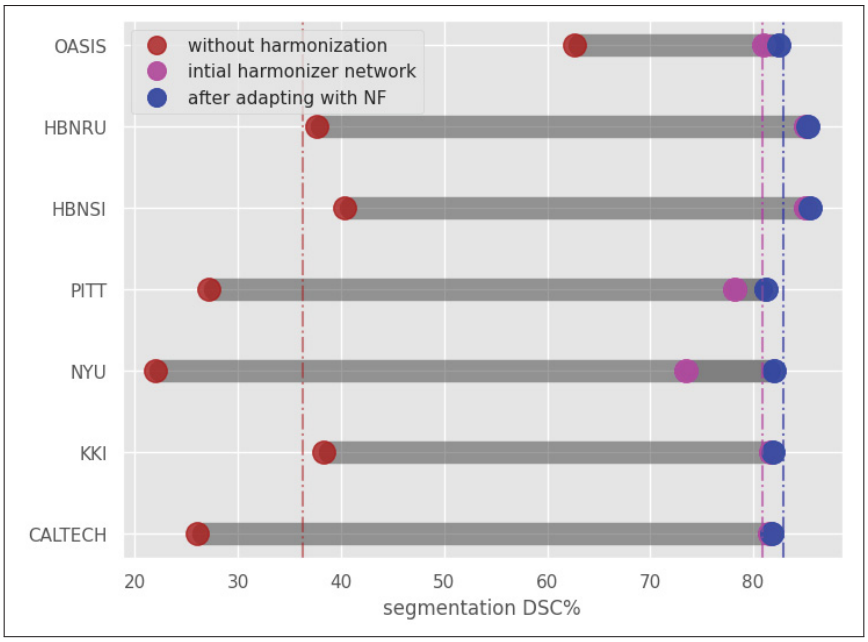



Figure 2.3 **Effect of each component of the Harmonizing Flows.** Particularly, we depict the improvement gained using the proposed pre-trained harmonizer network (~44 DSC%), and the adaptation using normalizing flows (~2.0 DSC%)

conducted in an unsupervised manner. Three different criteria to stop adapting the harmonizer network are explored. The first criterion is to stop in an iteration where the Shannon entropy of the target task predictions (segmentation in this case) reaches its minimum. According to (Wang *et al.*, 2021a), the Shannon entropy of the predictions of the target task is highly correlated with its performance. As it does not require any labeled data, it provides an *a priori* reliable stopping criterion for fine-tuning the harmonizer network. As a second alternative, we stopped the harmonizer network adaptation when the target *BPD* reaches the observed *BPD* on the source domain (which can be computed during training time using a validation set). As opposed to the first criterion, this criterion is task-agnostic and well-suited for unsupervised tasks or scenarios where entropy calculations are not applicable (i.e., regression problems). Finally, we directly used the target task (segmentation) performance and stopped the adaptation once it achieved the highest DSC score which we refer to as the *Oracle*. Please bear in mind that this criterion is impractical in real-world scenarios, as one may not have access to segmentation

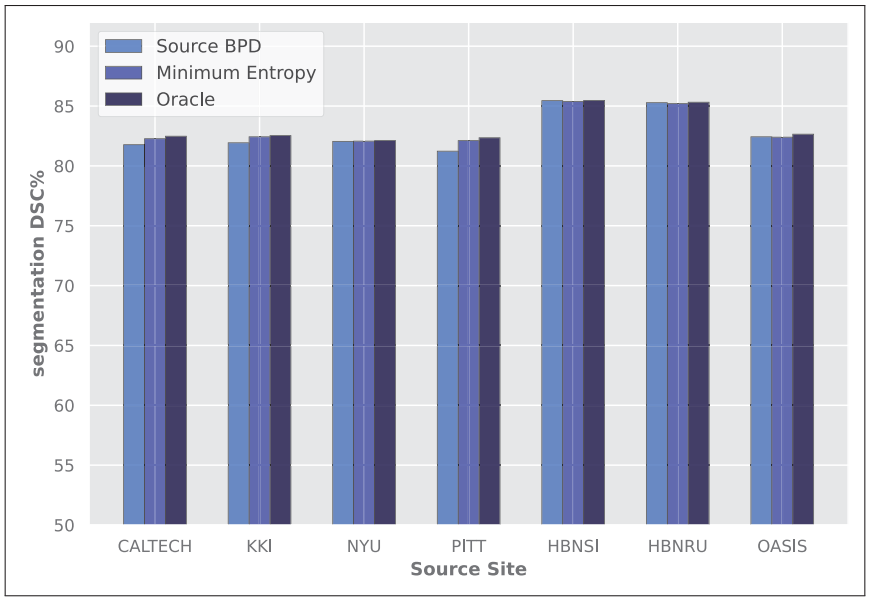



Figure 2.4 **The effect of different stopping criteria to stop the harmonizer network adaptation.** Oracle represents selecting the best iteration based on the performance of the target task (i.e., segmentation in this example), which serves as the upper bound. Both criteria, *source BPD* and minimum entropy, provide good stopping points, with a slight advantage of the minimum entropy criterion

labels[3], and it only intends to serve as an upper bound to show how much we can gain using a well-defined stopping criterion. As shown in Fig. 2.4, despite minimum entropy being a slightly better criterion compared to *source BPD*, both yield similar performances, with the *source BPD* providing a more general strategy, as it is not tailored to a task that requires probabilistic outputs. In summary, both stopping criteria prove to be viable options, as their outcomes closely resemble those of the *Oracle*.

Furthermore, in Fig. 2.5 we show, for a given scenario –harmonizing HBNSI dataset to NYU– the strong correlation between the segmentation metrics with both the segmentation prediction entropy and proximity to the source BPD. In particular, the evolution of the segmentation metrics is depicted in green (HD95) and blue (DSC) curves, with the values of the entropy of

[3] Please note that using segmentation, or other kind of labels, for the stopping criteria would make the model task-dependent.

the segmentation predictions in red, and the target BPD in purple. We can observe that the point for which these metrics are optimal (i.e., minimum entropy and target BPD matching source BPD), is actually close. Nevertheless, we advocate that, for more general use, looking at the source BPD should be a preferred option, as it does not depend on the target task.

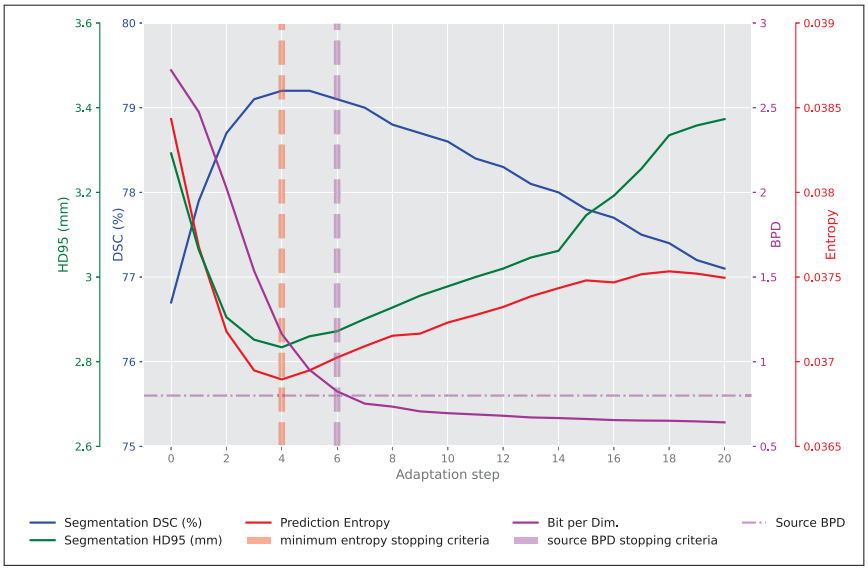
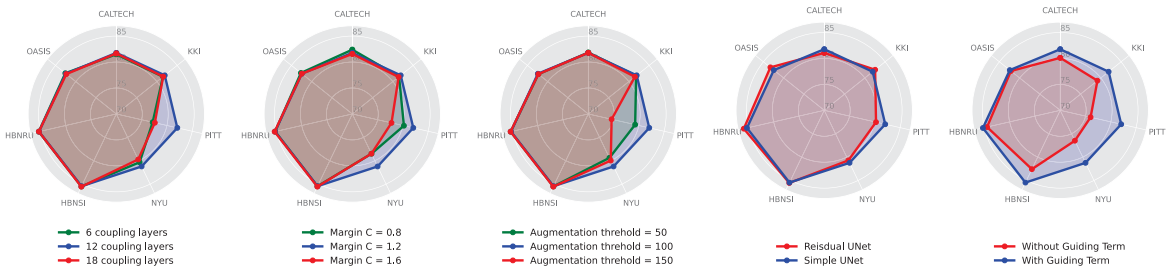



Figure 2.5 **Which is the best metric as stopping criteria?** This plot depicts different metrics during the adaptation of the harmonizer network (from HBNSI to NYU). Step zero corresponds to using the initial harmonizer network without adaptation. The vertical lines show the stopping time-points based on two proposed stopping criteria: minimum entropy of the predictions (*red*) and reaching source BPD *purple*)

**III-Ablation studies on components and hyper-parameters.** This section aims to empirically support the choices made in the proposed harmonization strategy.

**III.a-Flow Depth:**. We explored the effect of the number of coupling layers in the normalizing flow network in learning the distribution of the source domain. Particularly, we investigated using 6, 12, and 18 coupling layers in the normalizing flow network and its effect on the distribution learning capacity of the model and harmonization process. As can be seen in Fig. 2.6a, 12 coupling layers were the optimal choice and resulted in better harmonization. We believe that using 6 coupling layers does not fully capture the source domain distribution. On the other hand,




a) Number of used coupling layers in normalizing flow network.

b) Margin c for guiding the normalizing flow network.

c) Augmentation threshold to consider out of distribution.

d) Architecture of the harmonizer network.

e) Effect of the guiding term.

Figure 2.6 **Ablation study on architecture components and hyper-parameters of the harmonizing flow.** The best performance is obtained when the number of coupling layers in the normalizing flow is set to 12, the guiding term is used and the guiding margin $c$ to 1.2, the augmentation threshold is 100, and a simple UNet is chosen as the harmonizer network

using 18 coupling layers makes the training process harder, and it probably needs more data to be fully trained.

**III.b-Margin c :** In this section, we investigate the influence of the margin c on the guidance of the normalizing flow process. Our initial choice of 1.2 for c at the first level stemmed from observing an average source BPD of approximately 0.8 when the normalizing flow was trained without guidance. Consequently, we opted for a value of 1.2. To further investigate the effect of this value, we conducted additional experiments with margin values c set to 0.8 and 1.6. As illustrated in Fig. 2.6b, the initial choice yielded better performance in comparison to 0.8 and 1.6.

**III.c-Augmentation threshold:** Defining how much augmentation is enough for an augmented image to be considered as out of distribution to train the normalizing flow is an important step. This threshold, which is defined in terms of mean squared distance with original images, was first selected as 100, by visual inspection of the images. After the initial choice, we explored how adjusting this hyper-parameter affects the constraint on the normalizing flow network and

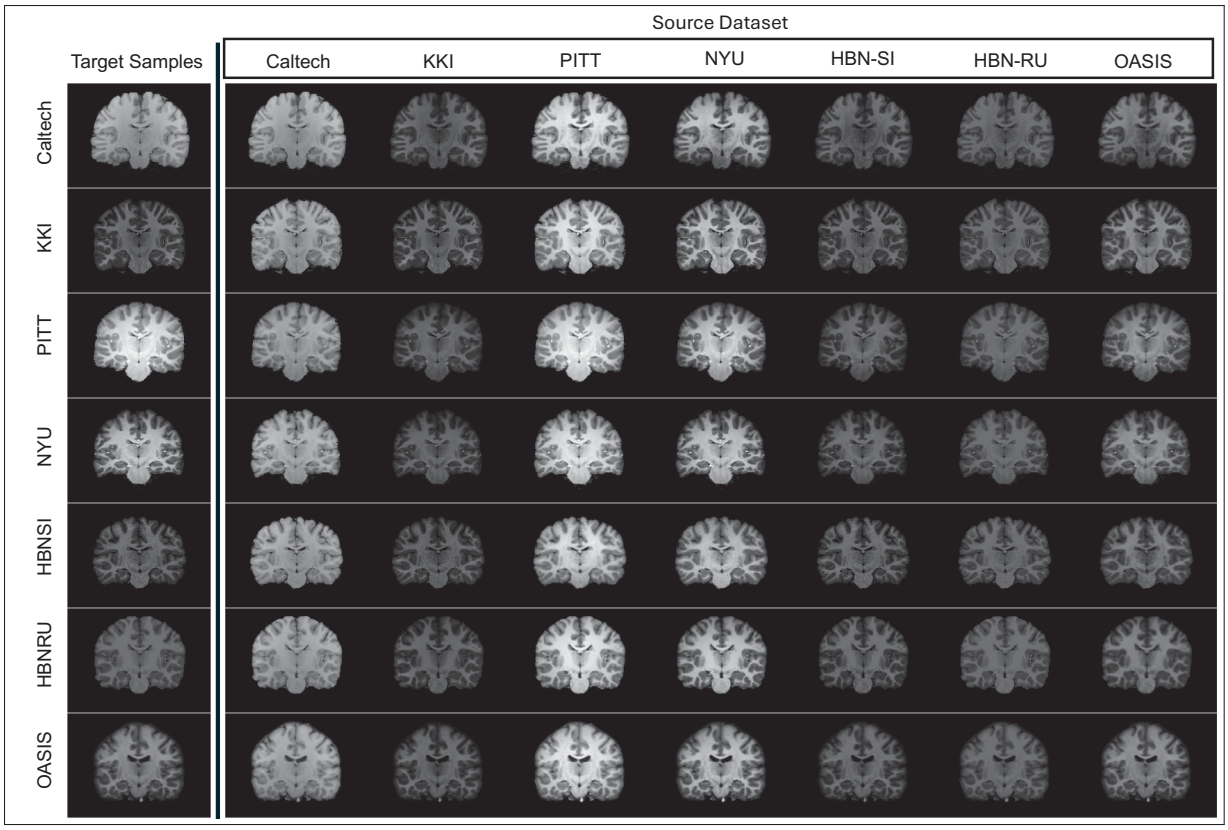



Figure 2.7 This figure showcases examples of harmonized images generated by the proposed method. The first column presents the sample images from various target domains. The subsequent columns display the images harmonized to specific source domains, as indicated at the top of each column. Therefore, each row maintains consistent anatomical structures, while each column shares the same visual characteristics

the harmonization process. We conducted experiments using augmentation thresholds of 50 and 150, whose results (Fig. 2.6c) demonstrate 100 to be an optimal choice for this hyperparameter.

**III.d-Harmonizer Network:** We investigated two U-shaped architectures for the harmonizer network. First, a conventional UNet architecture was employed. Subsequently, like the conference version of this paper (Beizaee *et al.*, 2023), we utilized a modified UNet to extract two separate sets of values. The final layer of the network ($\beta$) serves as a bias value, matching the input image's dimensions. Additionally, a scalar value $\alpha$, derived from the network's middle layer, acts as a scale parameter. In this way, the harmonizer's output can be expressed as $h_\theta(\mathbf{x}) = \alpha * \mathbf{x} + \beta$. As depicted in Fig. 2.6d, both options are effective, with a slight superiority

of the simple UNet. We believe that this marginal superiority is due to the greater degrees of freedom in simple UNet, which might be better for transferring complex distributions.

**III.e-Guiding Term:** In this section, we examine the impact of the guiding term by analyzing the effects of its removal from the training objective of the NF model (Eq. 2.6). As illustrated in Fig. 2.6e, incorporating the guiding term significantly enhances the results, underscoring its critical role in enabling the NF model to capture domain-specific characteristics of the source dataset. Conversely, omitting the guiding term often leads to suboptimal performance, adversely affecting the harmonization process. In many cases, the NF model trained without the guiding term produces results that are inferior even to those obtained with the initial harmonizer alone.

**Qualitative results.**

Fig. 2.7 showcases the instances of harmonized images using the proposed method across different source and target sites, for the adult brain MRIs. In particular, we randomly picked a sample from each site and then mapped it to different target sites. As can be seen, the harmonized samples in each column share the same visual characteristics, while on each row, the details of the harmonized images are preserved. These qualitative results illustrate that, regardless of the source or the target domains, the proposed method consistently produces reliable harmonized images, which is supported quantitatively by the comprehensive empirical validation conducted. Furthermore, visual examples depicted in Fig. 2.8 (neonatal brains) demonstrate the effectiveness of our approach in harmonizing inter-site images, regardless of the modality and plane used.

In Figure 2.9, we have visualized a harmonized sample from the target domain (CALTECH here) to the source domain (KKI here) using different harmonization methods. The harmonized image using our proposed method appears to have the closest visual characteristics with the source domain compared to other harmonization methods. Combat (Pomponio *et al.*, 2020) is excluded from this figure as it attempts to remove variations between domains and cannot transform the target domain to match the source domain.

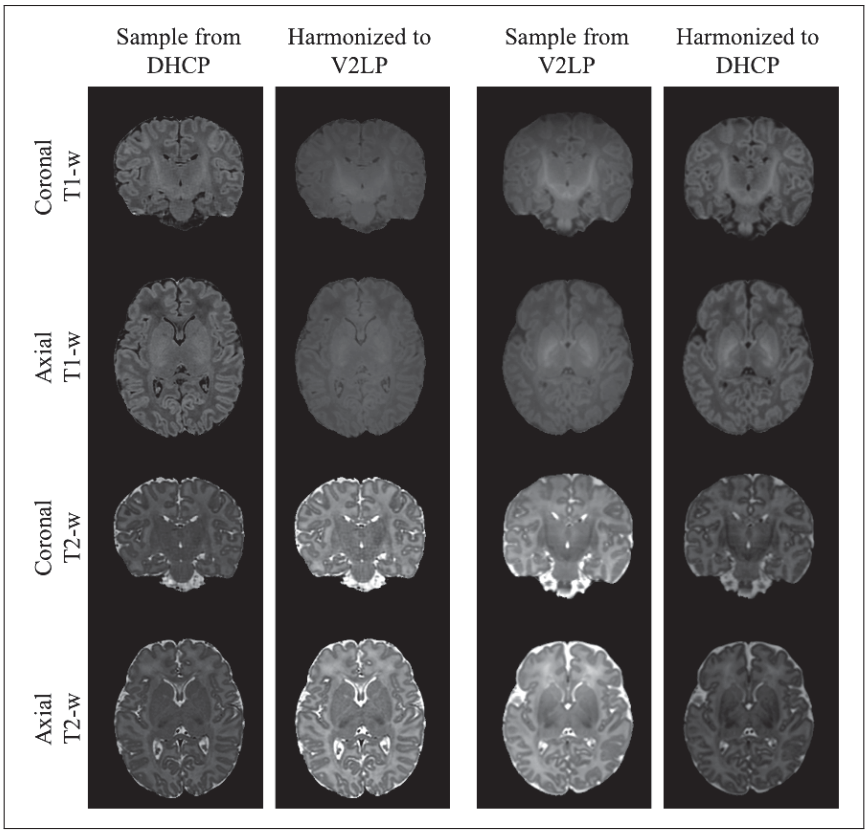



Figure 2.8 Visual examples of neonatal MRI brain images harmonized using the proposed method. The left section displays the coronal and axial planes for both modalities of a sample image from the DHCP dataset (target) alongside their harmonized counterparts to the V2LP dataset (source). The right section shows the reverse, with V2LP images harmonized to the DHCP dataset

#### 2.4.2.2 Performance on neonatal brain age estimation

In the previous section, we demonstrated the generalization capabilities of the proposed approach, by showing its superiority when evaluating the harmonized images in adults and neonatal brain MRI segmentation, and across multiple modalities. In this section, we will take one step further and evaluate the quality of the harmonization based on a different task, i.e., neonatal brain age estimation, which necessitates regression neural networks and the utilization of different 2D slices (specifically the axial view). As depicted in Table 2.4, our method also outperforms

Table 2.4 **Performance overview on the cross-site neonatal age estimation task.** Brain age estimation performance, in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE) metrics, across different harmonization approaches and modalities. To facilitate the strengths and weaknesses of different methods, we also indicate whether they are *source-free* ($\mathcal{SF}$), *task-agnostic* ($\mathcal{TA}$), and can handle *unknown-domains* ($\mathcal{UD}$), as well as the different strategy they fall in. The best results are highlighted in **bold**

| Method | $\mathcal{SF}$ | $\mathcal{TA}$ | $\mathcal{UD}$ | T1-w | | T2-w | |
|---|---|---|---|---|---|---|---|
| | | | | MAE | MSE | MAE | MSE |
| Baseline | – | – | – | 1.15 ±0.91 | 2.23 | 1.67 ±1.15 | 4.67 |
| Hist. matching (Nyúl *et al.*, 2000) | ✓ | ✓ | ✓ | 1.04 ±0.78 | 1.77 | 1.09 ±0.84 | 2.05 |
| Combat (Pomponio *et al.*, 2020) | ✗ | ✓ | ✗ | 1.33 ±0.84 | 2.49 | 1.28 ±0.82 | 2.33 |
| Cycle-GAN (Modanwal *et al.*, 2020) | ✗ | ✓ | ✗ | 1.05 ±0.71 | 1.62 | 0.89 ±0.63 | 1.24 |
| Style-transfer (Liu *et al.*, 2021) | ✗ | ✓ | ✗ | 1.75 ±1.17 | 5.15 | 0.96 ±0.67 | 1.41 |
| SSIMH (Guan *et al.*, 2022) | ✗ | ✓ | ✓ | 1.18 ±0.96 | 2.33 | 1.41 ±0.85 | 2.35 |
| Imunity (Cackowski *et al.*, 2023) | ✗ | ✓ | ✓ | 1.10 ±0.81 | 1.89 | 1.55 ±1.00 | 3.82 |
| BlindHarmony (Jeong *et al.*, 2023) | ✓ | ✓ | ✓ | 1.19 ±0.98 | 2.45 | 1.29 ±0.90 | 2.62 |
| AUE (Wang *et al.*, 2019) | ✓ | ✗ | ✓ | 1.32 ±1.01 | 2.91 | 1.4 ±1.01 | 3.51 |
| BigAug (Zhang *et al.*, 2020) | ✓ | ✗ | ✓ | 1.05 ±0.62 | 1.55 | 1.09 ±0.80 | 1.88 |
| **Harmonizing Flows** | ✓ | ✓ | ✓ | **1.01 ±0.69** | **1.51** | **0.82 ±0.68** | **1.21** |

all compared methods for both metrics and both modalities, highlighting its effectiveness for various tasks and applications, as well as on different 2D planes. Thus, based on the results of the segmentation and regression tasks, we can state that the proposed harmonization strategy leads to a significantly more flexible solution with substantial improvement gains.

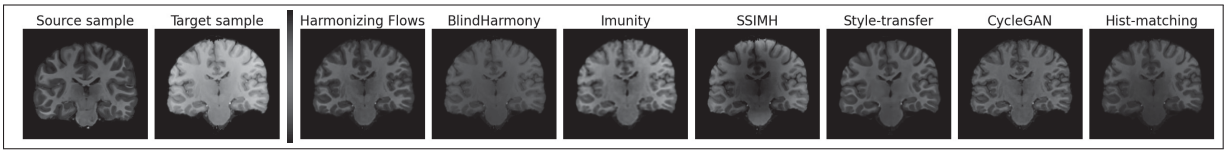



Figure 2.9 Example of harmonized images using different methods. The first image is the source domain sample (KKI here), the second one shows a sample from the target domain (CALTECH here), and the rest, shows the harmonized target sample using different methods

#### 2.4.2.3 A closer look at the harmonization performance

In previous sections, we assessed the quality of the harmonization based on the performance of different target tasks, i.e., adult and neonatal brain segmentation and neonatal brain age estimation. A natural question that arises is whether the harmonization performance can be quantitatively evaluated without requiring further labeled target tasks or *traveling subjects*. Indeed, recent literature (Parida *et al.*, 2024) in evaluating harmonization techniques has proposed the use of the Wasserstein distance, to measure the similarity between harmonized and source image intensity histograms. The reasoning behind using this kind of divergence is two-fold. First, standard harmonization metrics, such as mean absolute error, mean squared error, or peak signal-to-noise ratio, can provide high-quality measurements, but at the price of requiring paired harmonized data, i.e., *traveling subjects*. Second, while other metrics, such as structural similarity (SSIM), can help mitigate the need for paired data, they primarily compare structures at a higher level, potentially overlooking smaller artifacts or hallucinations introduced by generative models. In this section, we follow up on the recent work in (Parida *et al.*, 2024), and assess the harmonization performance of different strategies based on the Wasserstein distance between intensity histograms.

To this end, we first depict in Fig. 2.10 the histogram distributions of the harmonized MRIs from multiple target domains compared to the MRIs from the source site (the KKI dataset in this example), across all harmonization methods. Looking at these plots, we can observe that Histogram Matching, as well as our approach, yields to the closest intensity histogram distributions to the source domain (*in purple*), after harmonization. Furthermore, this behavior is consistent for all the target sites for both Histogram Matching and the proposed approach. Thus, from a pure harmonization standpoint, these plots tell us that Histogram Matching and our harmonization strategy are capable of mapping images from a target to a source site such that the harmonized intensity histograms almost match perfectly those of the source domain.

Nevertheless, perfect alignment between intensity histograms is not the ultimate goal. It should be noted that different individuals' brain structures (for example between a child and an old

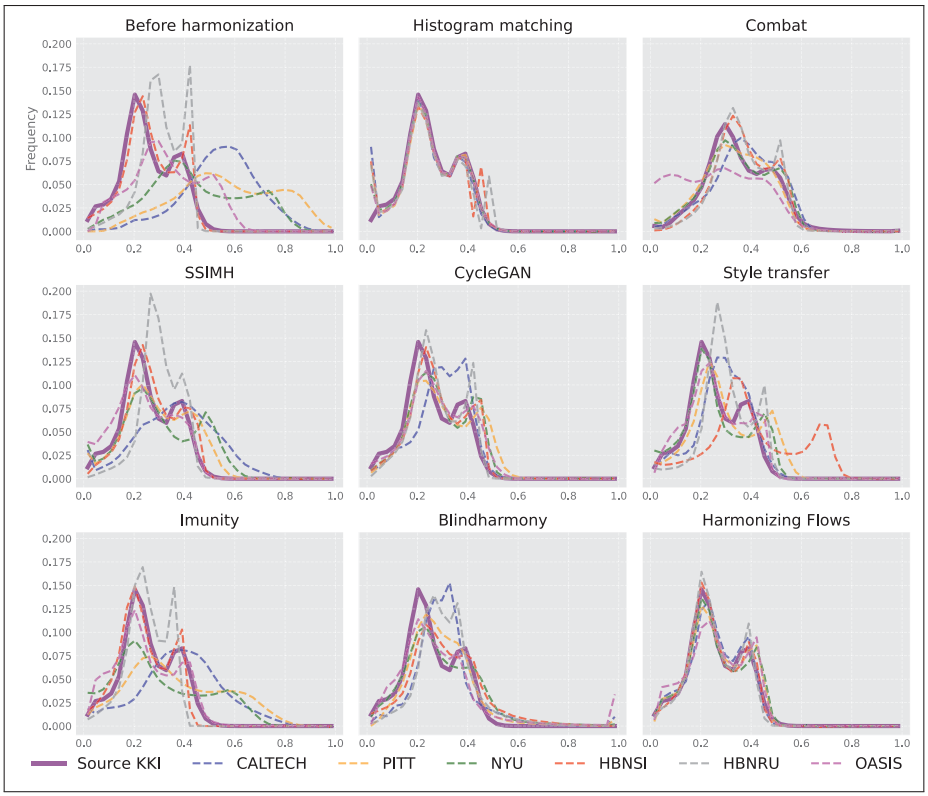



Figure 2.10 Histograms of the harmonized MRIs from multiple target datasets compared to the histogram of the source MRIs (KKI dataset, *in purple*) for all harmonization methods

person) necessitate different intensity histograms. Therefore, aligning intensity histograms is not an ideal solution. Also, we believe that the ability to enhance the performance of subsequent tasks, such as segmentation, classification, or brain age estimation in our context, which relies on harmonized images when dealing with multi-centric data, serve as a better indicator of the harmonization potential. However, if we analyze the performance on target tasks when employing harmonized data, we can easily observe that histogram matching indeed generates harmonized images that lead to poor performance on multiple target tasks. For example, in adult brain segmentation, resorting to histogram matching as a harmonization method results in an average DSC of 63.6, substantially lower than other methods which, a priori, had the highest misalignment between intensity histograms of the harmonized images with the source one in terms of WD (e.g, SSIMH or Imunity).

To better understand whether a correlation between the intensity histogram distance and the target task performance exists, we plot *task-performance vs. intensity-histogram-distance (WD)* in Figure 2.11. Indeed, we can observe from this figure that, in most cases, a closer intensity histogram (i.e., lower WD) corresponds to better performance in the target task (i.e., higher DSC in segmentation and lower MAE in brain age estimation). A method warranting further discussion is histogram matching, which shows a weak correlation, if any, between intensity histogram distance in terms of WD and target task performance. Its good performance in the WD metric is somehow expected, as it directly optimizes the histogram of intensities for MRI harmonization. Nevertheless, we believe that solely forcing the histograms to be close is not necessarily a good condition to yield usable harmonized images, as the spatial variation across intensities is not considered when matching distributions. Harmonized images depicted in Figure 2.9 could further illustrate this point visually. The remaining harmonization strategies, however, seem to be generally correlated, with approaches obtaining lower WD values also achieving better performances on the respective target tasks. In particular, our proposed approach always ranks among the top-three approaches in terms of WD, yielding the best performance in segmentation and age estimation tasks among compared harmonization methods. These findings are based on the metrics proposed in the recent work (Parida *et al.*, 2024), and may not hold for a more in-depth list of harmonization metrics. Ideally, while some other metrics could have been explored to evaluate the harmonization performance, some of the metrics used depend on *traveling-subjects*, which limits their applicability to many scenarios. We also stress that having an in-depth evaluation of the harmonization performance and assessing the best harmonization metric, is not within the scope of this work. In contrast, we believe that based on the results reported in this work, solely relying on the closeness of intensity histograms may not be sufficient, and evaluating the quality of the harmonization on target task indicators (e.g., DSC in segmentation or MAE in regression tasks) could serve as a more general manner to assess the harmonization quality.

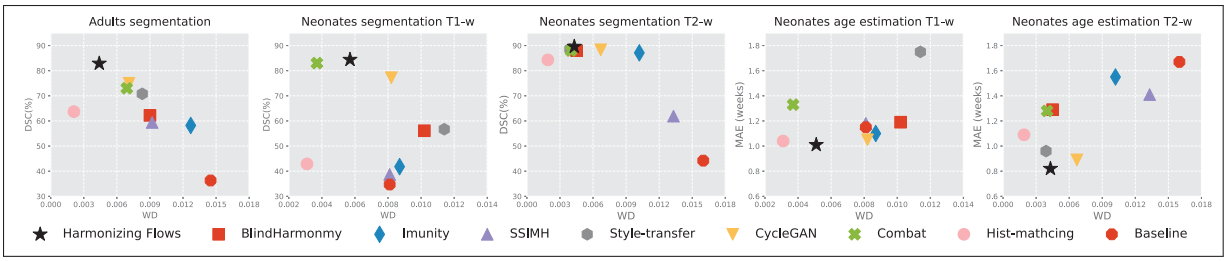



Figure 2.11 Segmentation (DSC%) and Age estimation (MAE) results vs. WD of intensity histograms for compared Harmonization methods

#### 2.4.2.4 Structural integrity.

One of the main concerns regarding learning-based MRI harmonization methods is anatomy hallucination (Cohen, Luck & Honari, 2018). Specifically, one issue that arises is whether training the normalizing flow model on a population different from the target domain with distinct anatomical variations could result in unintended changes to anatomical structures. As such, it is essential to implement strategies that ensure that structural integrity is preserved throughout the harmonization process. We have employed several techniques to mitigate this risk effectively.

First, to ensure robustness to structural variations, we applied a range of geometric augmentations while training the normalizing flow and harmonization model. These augmentations included resizing, scaling along the X, Y, and Z axes, shearing, perspective transformations, and deformations like piece-wise affine. These augmentations help the model focus on domain-characteristic features, such as brightness and contrast, while reducing sensitivity to structural differences. Second, during the pre-training of the harmonization network, we used the Structural Similarity Index (SSIM) loss, which, combined with geometric augmentation could help promote structural consistency and alignment between the input and output images. Finally, To further validate structural consistency, we specifically selected a segmentation task, which contains detailed regional contours, as a downstream evaluation. despite differences in age ranges between adult datasets, our method demonstrated superior performance compared to other harmonization techniques, providing empirical evidence of its structural fidelity.

It is important to note that evaluating the performance of harmonization in the presence of brain injuries, where the target domain contains injured brains and the source domain consists of normal brain MRIs, is not straightforward in our setting. Downstream tasks such as segmentation and age estimation, which are trained on normal brains, cannot be applied to abnormal brains, as these models are likely to fail. To demonstrate the structural integrity of the proposed harmonization method in the presence of brain injuries, we visually examined a neonatal brain MRI with atrophy and large ventricles from the V2LP domain before and after harmonization to the DHCP dataset. The results, depicted in Fig. 2.12 for both T1 and T2 modalities, reveal that the structures of the injured areas are preserved, highlighting the structural integrity of the proposed harmonization approach.

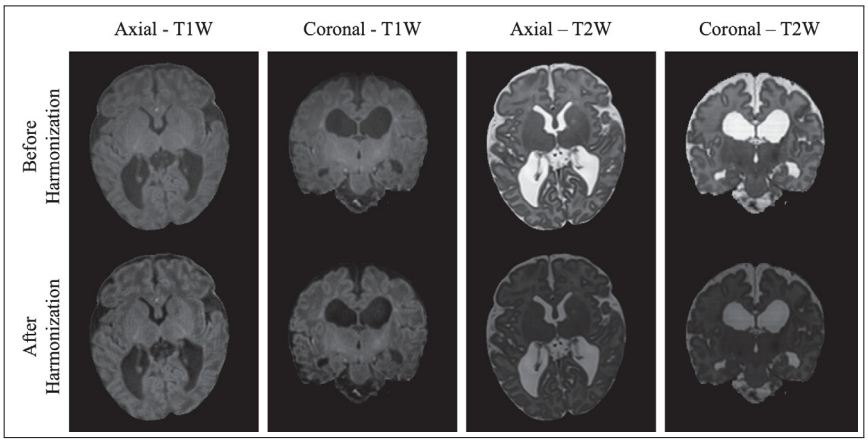



Figure 2.12 Visualization of harmonization applied to a sample of neonatal injured (atrophy and large ventricles) brain MRI from the V2LP dataset, harmonized to the DHCP dataset. The results demonstrate that the model preserves the structural integrity of the input while performing harmonization, even in the presence of injury-related anomalies

#### 2.4.2.5 Friedman Ranking

To fairly compare the performance of the different harmonization methods across various metrics and tasks, we resort to the Friedman Rank (Friedman, 1937, 1940), which has been employed

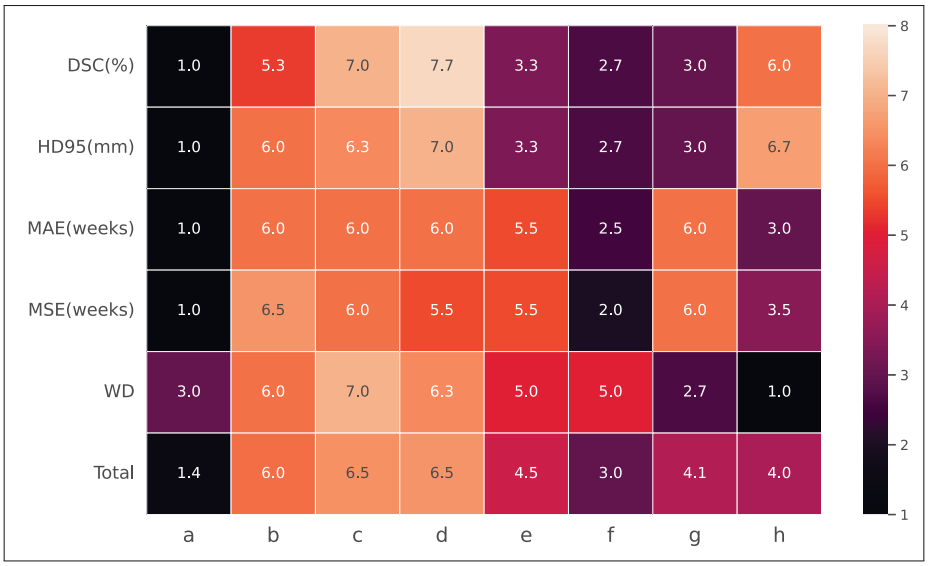



Figure 2.13 Friedman Rank for the compared harmonization methods: a) Harmonizing Flows (Ours), b) BlindHarmony, c) Imunity, d) SSIMH, e) Style-transfer, f) CycleGAN, g) Combat, h) Hist-matching

for this purpose in the literature (Wang *et al.*, 2022; Murugesan *et al.*, 2025). The **Friedman Rank** is defined as:

$$\text{rank} = \frac{1}{S_m} \sum \text{rank}_i \tag{2.8}$$

where $S_m$ is the number of evaluation settings and $\text{rank}_i$ is the rank of a method in the $i$-th setting. Thus, the lower the rank obtained by an approach, the better the method is. In our scenario, we have 13 different settings: DSC and HD95 in 3 segmentation tasks, MAE and MSE in 2 regression problems, and 3 WD values in harmonization (one for adults and two for neonatal brains). The results from the Friedman Ranking across all analyzed methods are depicted in Figure 2.13. As it can clearly observed, our proposed method (*last column*) achieves the best Friedman rank among all compared harmonization methods, demonstrating their overall superiority across different scenarios.

## 2.5 Conclusion

In this work, we proposed a novel harmonization method that leverages Normalizing Flows to guide the adaptation of a harmonizer network. Our approach is source-free, task-agnostic, and works with unseen domains. These characteristics make our model applicable in real-life problems where the source domain is not accessible during adaptation, target domains are unknown at training time and harmonization is task-independent. Furthermore, another advantage of our method over the existing approaches is that it only requires images from one source domain, and one modality, at the training time.

Through extensive comparisons with other harmonization methods, as well as test-time domain adaptation and generalization approaches, our method consistently proved its superiority in multiple medical image problems, yet relaxing the strong assumptions made by existing harmonization strategies. Furthermore, we validated the scalability of our proposed method by evaluating it on a different population (neonates), different modalities (T1-weighted and T2-weighted MRIs), and different tasks (neonatal brain age estimation and segmentation). The results demonstrated comparable performance across diverse populations and modalities, highlighting the robustness and versatility of our approach. It is worth noting that while this paper applies harmonization to 2D brain MRI slices, the method can be readily extended to 3D by replacing 2D convolutional layers with 3D ones. However, this extension would require a larger dataset and more computational resources due to the increased number of parameters. Additionally, although no visible inconsistencies between harmonized slices were observed in our experiments, averaging harmonization results across all 2D planes (i.e., axial, coronal, and sagittal) could further mitigate potential inconsistencies. Qualitative results further supported the reliability and effectiveness of our method, illustrating consistent and reliable image-to-image mappings across different target domains. Last, even compared to recent test-time adaptation strategies, empirical results suggest that the proposed method is a powerful alternative to deal with the presence of domain drifts, more particularly for MRI multi-site harmonization.

In conclusion, our proposed harmonization method offers a promising solution for addressing distributional shifts in medical image analysis, paving the way for improved performance and generalizability across diverse datasets, and enabling the use of large-scale multi-centric studies.

**Limitations:** Although the proposed harmonization method demonstrates promising potential, there are a few considerations. First of all, it requires additional inference time for test-time adaptation, which may impact efficiency in certain time-sensitive applications. Moreover, while efforts have been made to prevent the normalizing flow model from capturing biases or anatomical structures in the source domain, it may still inadvertently learn some biases, such as image quality variations. Furthermore, our approach assumes that the source and target domains have close resolutions, which may limit its flexibility when applied to datasets with varying resolutions.

# CHAPTER 3

# MAD-AD: MASKED DIFFUSION FOR UNSUPERVISED BRAIN ANOMALY DETECTION

Farzad Beizaee[1,2] , Gregory Lodygensky[2,3] , Christian Desrosiers[1] , José Dolz[1]

[1] ÉTS Montreal, Canada
[2] CHU Sainte-Justine, University of Montreal, Canada
[3] University of Montreal, Canada

Article accepted in the "Information Processing in Medical Imaging (IPMI)", February 2025

## 3.1 Introduction

The accurate detection and localization of brain anomalies in medical images, particularly in Magnetic Resonance Imaging (MRI) data, is paramount to diagnosing and understanding neurological injuries and pathologies. However, the complexity of brain structures and the scarcity of labeled abnormal data present significant challenges in developing robust and generalizable solutions. Traditionally, brain anomaly detection has been framed as a supervised learning task, which aims at identifying well-defined pathologies such as brain tumor (Havaei *et al.*, 2017; Kamnitsas *et al.*, 2018; Sinha & Dolz, 2020), atrophy (Pagnozzi, Fripp & Rose, 2019) or white matter hyper-intensities (Kervadec *et al.*, 2021; Kuijf *et al.*, 2019), among many others. Nevertheless, casting anomaly detection as a supervised problem introduces an inherent bias towards the targeted lesions, limiting the scope of detectable pathologies. Moreover, collecting large amounts of annotated samples encompassing the entire spectrum of potential brain abnormalities is expensive and impractical for novel structures or rare abnormal patterns.

Unsupervised anomaly detection (UAD), which involves modeling the distribution of normal data and identifying deviations as anomalies, has gained attention as a promising alternative (Behrendt *et al.*, 2022; Bercea *et al.*, 2024b; Silva-Rodríguez *et al.*, 2022; Zimmerer *et al.*, 2019). Conventional unsupervised methods, such as autoencoders (Rumelhart, Hinton & Williams, 1986) and generative adversarial networks (GANs) (Goodfellow *et al.*, 2014), attempt to reconstruct normal anatomical structures and flag areas with high reconstruction errors as anomalies. Despite

their potential, these approaches suffer from notable limitations. Autoencoders often fail to capture the fine-grained details of normal anatomy, whereas GANs are prone to mode collapse and instability during training. Moreover, these models frequently reconstruct anomalies as part of normal structures, reducing their reliability in clinical applications.

Recent advances in diffusion models (Ho *et al.*, 2020b; Kingma *et al.*, 2021; Sohl-Dickstein *et al.*, 2015; Song *et al.*, 2021a) have opened new avenues in generative modeling. Such models (Sohl-Dickstein *et al.*, 2015) leverage a stochastic process to gradually corrupt data and learn to reverse this process, enabling them to model complex data distributions with remarkable precision. Their success in generating high-quality images and their ability to capture intricate patterns in the data have prompted researchers to explore their use for anomaly detection (Graham *et al.*, 2023; Lu *et al.*, 2023; Yan *et al.*, 2023; Liang *et al.*, 2023; Naval Marimont *et al.*, 2024). While these methods have improved the accuracy of anomaly detection, their application to brain images introduces several challenges. Firstly, the forward diffusion process can cause a loss of distinctive features across brain regions, especially when the number of steps is large. This loss may compromise the model's ability to differentiate between normal and anomalous brain regions. Also, reducing the number of forward diffusion steps introduces the risk of an *"identity shortcut"* problem. In this problem, the model can easily recover the fine details of the input image, resulting in anomalous regions being preserved in the reconstruction. This is a significant concern in brain anomaly detection, where subtle but critical deviations such as tumors or lesions may be overlooked due to this shortcut behavior. Another issue arises from the indiscriminate application of forward and reverse diffusion across the entire brain image. This approach can hinder the model's ability to effectively reconstruct normal brain patterns.

To address these limitations, we propose MAD-AD, a Masked Diffusion for brain anomaly detection with the following key contributions. First, we leverage latent diffusion models to treat anomalies as partial noise in the latent space, enabling their effective restoration through the denoising process. Our method also removes the reliance on forward diffusion steps during inference, thereby preventing the loss of critical visual details and enabling highly accurate reconstructions of the underlying normal appearance. This is accomplished by masking the

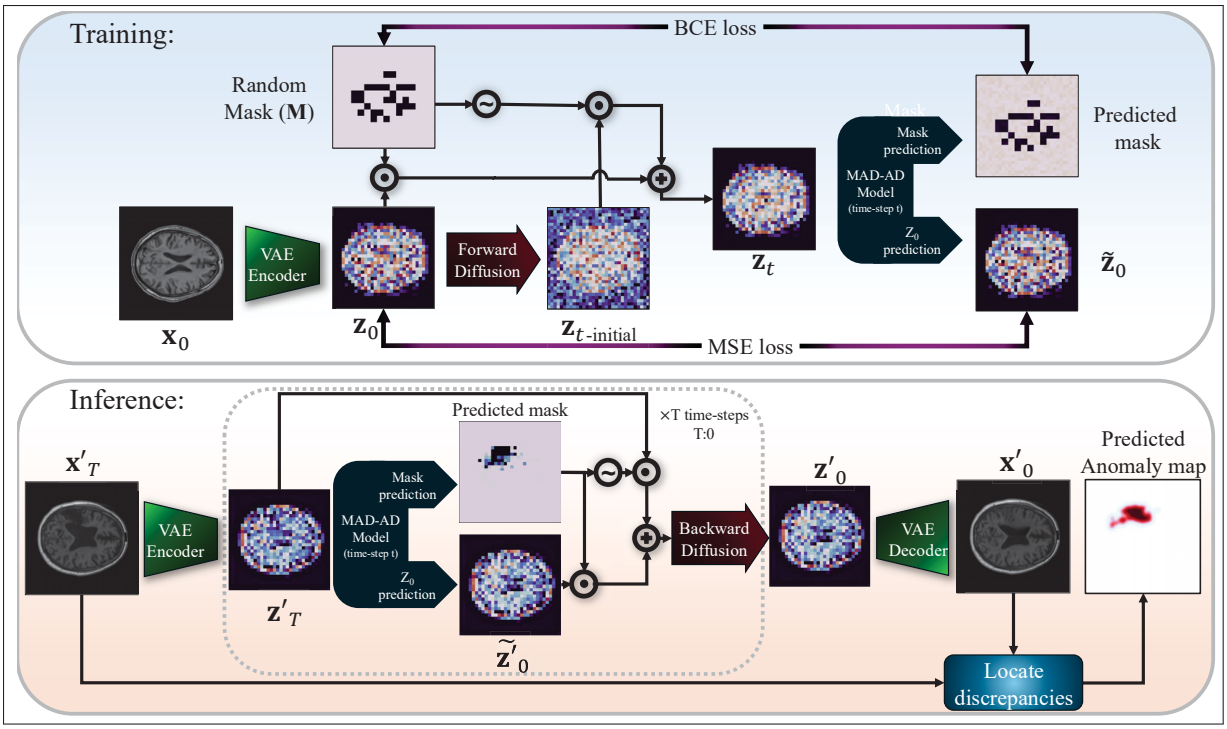



Figure 3.1 **Overview of the proposed method.** During *training*, normal samples are encoded into latent space. A binary mask and a time-step t are applied, and non-masked regions undergo forward diffusion to produce $z_t$. The model is then trained to predict $z_0$ and the incorporated mask for forward diffusion. At *inference*, the model undergoes a selective reverse process using the predicted mask at each step

forward diffusion process and training the model to reverse it effectively. Furthermore, we incorporate a mask-prediction module into the diffusion framework, allowing the prediction of the incorporated mask in the diffusion process. This approach ensures the selective correction of anomalous regions while preserving normal regions intact, ultimately delivering more precise and reliable anomaly detection results. The overview of our method is depicted in Fig. 3.1.

## 3.2 Related works

Recent approaches for unsupervised anomaly detection (UAD) in brain MRI can mainly be divided in three categories: methods based on different variants of autoencoders (AEs), those using generative adversarial networks (GANs) and the ones based on diffusion models.

**AE-based methods.** Approaches in this category train an autoencoder on normal data to accurately reconstruct input images. At inference, the reconstruction error measured at each pixel is used to localize anomalies. Different networks have been explored for the reconstruction, including standard autoencoders (AE) (Atlason *et al.*, 2019; Baur *et al.*, 2021a), variational autoencoders (VAEs) (Baur *et al.*, 2021a; Silva-Rodríguez *et al.*, 2022; Zimmerer *et al.*, 2019) and denoising autoencoders (DAEs) (Kascenas, Pugeault & O'Neil, 2022). A common issue with these methods is their propensity to overfit the training data, leading to a poor generalization on unseen data. Furthermore, they are prone to blurry reconstructions, struggling to accurately distinguish subtle anomalies from normal variations, especially when relying solely on reconstruction error as a measure of abnormality.

**GAN-based methods.** These approaches employ an adversarial learning strategy where a generator and a discriminator are jointly trained on healthy subject images to learn a latent representation of normal variability. AnoGAN (Schlegl *et al.*, 2017) measures anomaly scores based on a combination of reconstruction error and distance in the latent space. f-AnoGAN (Schlegl *et al.*, 2019a) improves upon this work by incorporating an additional feature-level reconstruction strategy, yielding a more precise localization of anomalies. The work in (Baur *et al.*, 2020a) uses a style transfer method based on CycleGANs to map real MR images of healthy brains to synthetic ones, and vice versa. Anomalies are then detected by comparing input images to their reconstruction. While the ability of GANs to generate high-quality images can translate in a more detailed delineation of anomalies, they are also prone to training instability and are often sensitive to hyperparameter choices.

**Diffusion-based method.** Diffusion models have gained significant attention in computer vision for their ability to generate high-fidelity images (Croitoru *et al.*, 2023). Recently, these models have also shown promise in various medical image analysis tasks including UAD (Behrendt *et al.*, 2024b; Iqbal *et al.*, 2023; Behrendt *et al.*, 2024c; Bercea *et al.*, 2024a; Liang *et al.*, 2024b; Marimont *et al.*, 2024; Naval Marimont *et al.*, 2024; Wyatt *et al.*, 2022b). A prominent diffusion-based method for UAD in medical images, AnoDDPM (Wyatt *et al.*, 2022b) utilizes a partial diffusion strategy, adding noise to an image up to a specific timestep and then recovering

the original image with a reverse diffusion process. This method has shown success in detecting anomalies in brain MRI and other domains. PDDPM (Behrendt *et al.*, 2024b) instead applies the diffusion process in a patch-wise manner, aiming to improve the understanding of local image context and achieve better anatomical coherence in the reconstruction. This method divides the image into overlapping patches and reconstructs each patch while considering its unperturbed surroundings. CDDPM (Behrendt *et al.*, 2024c) generates multiple reconstructions via the reverse diffusion process and pinpoints anomalies by examining the distribution of these reconstructions with the Mahalanobis distance, subsequently labeling outliers as anomalies. MDDPM(Iqbal *et al.*, 2023) incorporates masking-based regularization, applied on both image patches and in the frequency domain, to enhance unsupervised anomaly detection. AutoDDPM (Bercea *et al.*, 2023a) incorporates automatic masking, stitching, and resampling techniques within the DDPM framework to enhance its robustness and accuracy in anomaly detection. This approach also addresses the challenge of selecting an appropriate noise level for detecting lesions of various sizes. However, the diffusion-based UAD models mentioned above rely heavily on a forward diffusion process that inherently results in information loss. Consequently, these methods often fail to accurately reconstruct the original healthy brain structures, leading to false-positive detections where normal regions are incorrectly identified as anomalous. This issue is particularly prominent in brain anomaly detection tasks, as brain structures, especially cortical regions, vary uniquely across individuals, thereby increasing the difficulty of accurately recovering normal anatomical variations.

A recently proposed method, DISYRE (Marimont *et al.*, 2024; Naval Marimont *et al.*, 2024), uses a diffusion-like pipeline to train a model to restore images that have been corrupted with synthetic anomalies. Anomalies in a new image are detected based on the model's ability to restore the image to a healthy state. A key limitation of this method is that the synthetic anomalies may not encompass all types of real-world anomalies, limiting its generalization ability. THOR (Bercea *et al.*, 2024a) integrates implicit guidance into the DDPM's denoising process using intermediate masks to preserve the integrity of healthy tissue details. It aims to ensure a faithful reconstruction of the original image in areas unaffected by pathology, minimizing false positives.

However, since these intermediate masks are determined based on the perceptual differences between input images and their reconstruction at each step, the model may struggle to detect subtle or small anomalies, as they might be masked out due to their minimal differences with the input image. Additionally, reconstruction errors may occur due to the loss of details during the forward process, with normal regions not getting masked due to their high perceptual differences. Inspired by diffusion-based models, IterMask$^2$ (Liang *et al.*, 2024b) incorporates an iterative spatial mask refinement process and frequency masking to enhance UAD performance. This strategy minimizes information loss in normal areas by iteratively shrinking a spatial mask, starting from the whole brain towards the anomaly. Although the model performs well in detecting hypo- or hyper-intense areas, it can fail to localize structural anomalies such as atrophy or enlarged ventricles as their reconstruction is conditioned on structural information from high-frequency image components which can be recovered by the model.

## 3.3 Method

### 3.3.1 Modeling the normal feature space

We resort to diffusion models for learning the space of normal data and reconstructing the normal counterpart of anomalous regions. Denoising Diffusion Probabilistic Models (DDPMs) (Ho *et al.*, 2020b) learn a data distribution by gradually adding noise to the data (i.e., forward process) and then training a model to reverse this process. While DDPMs are highly effective at generating high-quality images, there are certain limitations when using them directly for detecting anomalous regions. Firstly, the number of steps in the forward diffusion process can have a considerable impact on the performance. If this number is too large, semantic information of the brain structure can be lost, resulting in an uncorrelated brain reconstruction and the incorrect detection of normal regions as abnormal. On the other hand, if not enough steps are used, the model can too easily recover the fine details in the image. As a result, abnormal regions will incorrectly be detected as normal. Moreover, as normal patches are also affected by noise, they cannot be fully exploited to reconstruct abnormal regions.

To overcome the aforementioned limitations, we propose to incorporate a random masking strategy in the diffusion model and modify the reverse process so that the diffusion model can selectively alter anomalous parts of an image, while keeping the normal regions untouched. Following (Rombach *et al.*, 2022; Peebles & Xie, 2023), we employ a diffusion model operating in the *latent* space. This has two important advantages. First, whereas adding Gaussian noise directly on the image yields corruptions that have no meaningful structure, injecting this noise on latent features and then reconstructing these noisy features results in more complex corruptions that better represent real anomalies in brain MRI. Moreover, this also mitigates the "identity shortcut" problem, enhances computational efficiency, and improves stability, particularly with limited training data.

Let $\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ be the training set consisting exclusively of normal images $\boldsymbol{x}^{(i)} \in \mathbb{R}^{H\times W\times C}$, where $H$, $W$, and $C$ correspond to the image height, width, and number of channels, respectively. We employ a pre-trained variational autoencoder (Rombach *et al.*, 2022), which is adapted and fine-tuned for medical images. This model can encode high-dimensional image data into a compact latent representation and reconstruct this data from the latent space while preserving essential structural and semantic information. Denoting the encoder network as $\mathrm{V}_{E,\phi}$, an input image $\boldsymbol{x}^{(i)}$ is mapped to its latent space representation $\boldsymbol{z}^{(i)} = \mathrm{V}_{E,\phi}(\boldsymbol{x}^{(i)})$, where $\boldsymbol{z}^{(i)} \in \mathbb{R}^{H'\times W'\times C'}$.

**Random masking.** To incorporate random masking into the forward diffusion process, given the latent features of an input normal sample $z_0 \sim p\,(z_0)$, we spatially divide $z_0$ into non-overlapping patches defined by a random mask $M \in [0,1]^{H\times W}$. The forward Markov diffusion process to generate samples $z_t$ gradually applies noise to the non-masked patches of sample $z_0$ for $t$ time steps, where $t \in [1, T]$. Following (Ho *et al.*, 2020b), the forward noising process in the latent space with masking can be characterized as:

$$z_t = \left(\sqrt{1-\beta_t}z_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}\right) \odot M + z_0 \odot (1 - M), \tag{3.1}$$

where $z_t$ is the partially diffused image at step $t$, $\epsilon \sim \mathcal{N}(0, I)$ is the sampled Gaussian noise and $\beta_t$ is the noise schedule at step $t$, which controls the amount of noise added at each step. Using

the reparameterization trick, $z_t$ can be obtained implicitly using the following equation:

$$z_t = \left(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}\right) \odot M + z_0 \odot (1-M), \tag{3.2}$$

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$. The reverse process aims to recover the original data $z_0$ by gradually removing the noise. This process is modeled as a learned distribution that reverses the forward noising steps. Given the masked sample $z_t$ at step $t$ and mask $M$ at spatial location $k$, the reverse process can be modeled as follows:

$$p(z_{t-1}^k | z_t^k) = \begin{cases} \mathcal{N}(z_{t-1}^k; \mu_\theta(z_t^k, t),\ \beta_t \mathbf{I}), & \text{if } M^k = 1 \\ z_t^k, & \text{otherwise}; \end{cases} \tag{3.3}$$

In this equation, $\mu_\theta(z_t, t)$ is a trainable function, which can be reparameterized as a predicted noise $\epsilon$ or a predicted clean image $z_0$. Due to the incorporated random masking strategy, we prefer the latter one for simplicity. Therefore, $\mu_\theta(z_t, t)$ can be formally expressed as:

$$\mu_\theta(z_t, t) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} f_{\theta, z_0}(z_t, t) + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} z_t, \tag{3.4}$$

where $f_{\theta,z_0}(z_t, t)$ is a function that predicts $\tilde{z}_0$ at time step t, given $z_t$.

**Mask prediction.** By parameterizing $f_\theta$ as a neural network, the model can be trained using a simple mean-square error loss between $z_0$ and the predicted clean image. Moreover, in Eq. (3.3), we assumed that the mask $M$ is available in the reverse process. However, this assumption is unrealistic since the mask used in diffusing the image, which contains the location of anomalous regions, is not accessible at inference. Therefore, we include an additional head $f_{\theta,M}$ to the diffusion model that predicts the mask used in the forward diffusion. This can be achieved by applying a binary cross-entropy ($\mathcal{L}_{BCE}$) loss between the predicted mask from this head and a randomly sampled mask used during partial diffusion in training. The final training objective of

our model is defined by:

$$\min_{\theta} \ \mathbb{E}_{z_0 \sim q(z_0), \epsilon, t, M} \left[ \left\| z_0 - f_{\theta, z_0}\left(z_t, t\right) \right\|_2^2 + \lambda \mathcal{L}_{BCE}\big(M,\ f_{\theta, M}(z_t, t)\big) \right], \tag{3.5}$$

where $\lambda$ is a hyper-parameter that balances the contributions of the two terms.

### 3.3.2 Recovering normal images

During inference, the goal is to recover a normal version of an abnormal brain image, where anomalous regions are replaced with their normal counterpart while normal areas remain unchanged. As previously discussed, a pre-trained VAE, $V(\cdot)$, is employed to project the image into a latent space where the data follows a normal distribution. In this space, abnormal brain regions can be interpreted as normal noise, as they fall outside the learned normal distribution of the model. These abnormal areas can also be considered as non-masked regions through the forward diffusion process using a mask that points out anomalous regions. Consequently, the proposed method incorporates all the necessary components to first predict the location of anomalies using the mask prediction head and then progressively denoise these regions to reconstruct their normal counterpart. Finally, by comparing the input image with its corrected version, anomalies can be accurately localized. The following section provides a detailed explanation of the sampling process in the MAD-AD model during inference.

Let $\mathcal{X}' = \{\boldsymbol{x}'^{(i)}\}_{i=1}^{N'}$ denote the test set at inference time, which consists of samples with potential anomalies. We first map these images into the latent space using $\mathrm{V}_{E,\phi}$. As explained before, we treat the latent space of an anomalous image as step $T$ of the masked forward diffusion process applied on its normal counterpart, i.e., $z'_T = \mathrm{V}_{E,\phi}(\boldsymbol{x}'_T)$. By predicting the mask that corresponds to the anomaly location and the reconstructed $\tilde{z}'_0$ at each time-step $t$, using Eq. (3.3), we can progressively correct the anomaly regions and obtain the normal counterpart ($z'_T \rightarrow z'_0$) while preserving fine details of the normal regions.

Nevertheless, one drawback of sampling with DDPM is that it requires many reverse sampling steps to obtain the normal version. Therefore, we instead opted for DDIM (Song *et al.*, 2021a)

which, by reducing the stochasticity of DDPM, makes the reverse process more deterministic and requires fewer sampling steps. Consequently, we modify the reverse process of DDIM for the MAD-AD model as:

$$\begin{aligned} z'_{t-1} = \underbrace{B\Big(f_{\theta,M}(z'_t)\Big)}_{\text{“predicted mask”}} \Big(\sqrt{\bar{\alpha}_{t-1}}\ \underbrace{f_{\theta,z_0}(z'_t)}_{\text{“predicted } \tilde{z}'_0\text{”}} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}\tilde{\boldsymbol{\epsilon}}_t(z'_t)}_{\text{“direction pointing to } z'_t\text{”}} + \sigma_t \boldsymbol{\epsilon}'_t\Big) \\ + \Big(1 - B\big(f_{\theta,M}(z'_t)\big)\Big) \cdot z'_t \end{aligned} \tag{3.6}$$

where $B(.)$ is a binarization function, $\boldsymbol{\epsilon}'_t$ is random normal noise, $\sigma_t$ is a hyper-parameter that controls the stochasticity of reverse process, and $\tilde{\boldsymbol{\epsilon}}$ is the predicted noise calculated based on the predicted $\tilde{z}_0$ and $z_t$ as follows:

$$\tilde{\boldsymbol{\epsilon}}_t = \frac{z'_t \cdot f_{\theta,z'_0}(z'_t)}{\sqrt{1-\bar{\alpha}_t}} \tag{3.7}$$

As mentioned above, $\sigma_t$ controls the noise level and stochasticity of the sampling process in DDIMs. Specifically, $\sigma_t = 0$ makes the model deterministic, while $\sigma_t > 0$ introduces stochasticity. For $\sigma_t = 1$, the model behaves like a DDPM, where the sampling process involves full stochasticity with noise added at each step. While having a fully deterministic model can be desirable for UAD applications, introducing a bit of noise to the non-masked (anomalous) regions helps bring the distribution closer to normal. This makes it easier for the model to recover the normal variation of the input. Therefore, we propose to use an in-between value of $\sigma_t = 0.5$. A qualitative example of the reverse process in MAD-AD is depicted in Figure 3.2.

### 3.3.3 Anomaly localization

Equation 3.6 enables a correcting trajectory from $z'_T$ to $z'_0$, resulting in generating high-quality normal variation of the anomalous image in fewer steps. To accurately localize anomalies, we used the discrepancy between the input image and its reconstructed normal counterpart. More concretely, using the “normal” latent embedding $\tilde{z}'_0$, we generated a reconstructed normal sample in the image-space as $\tilde{x}'_0 = \mathcal{V}_{D,\phi}(\tilde{z}'_0)$, where $\mathcal{V}_{D,\phi}$ is the pre-trained VAE decoder. The predicted

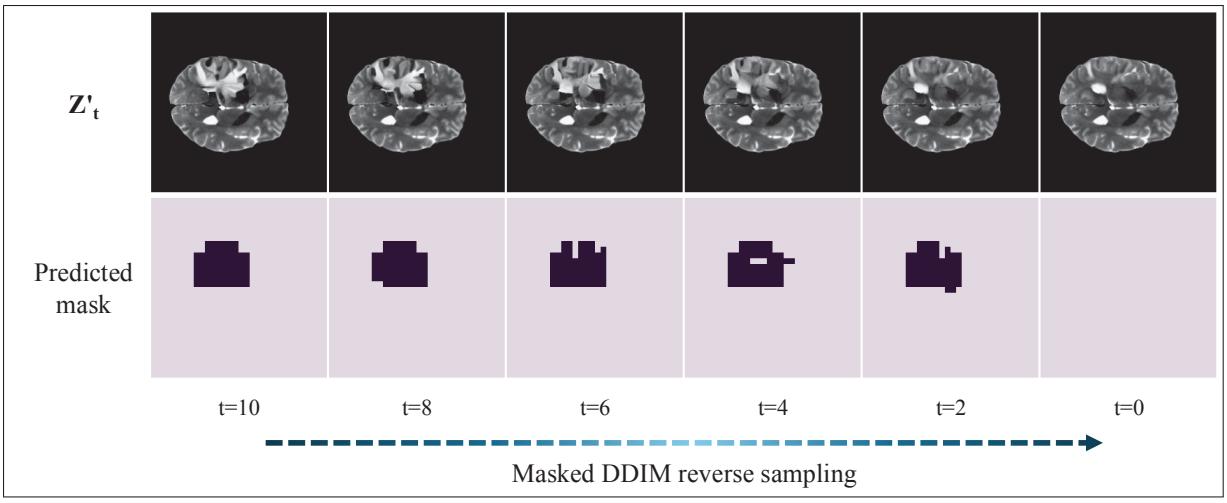



Figure 3.2 **Visual example of the reverse process.** Both the predicted mask and the decoded latent representation of the intermediate reverse step ($z'_t$) at multiple time steps are depicted to highlight the masked reverse sampling in MAD-AD

anomaly map is then given by:

$$\boldsymbol{a} = G * \min(\|\tilde{\boldsymbol{x}}'_0 - \boldsymbol{x}'_0\|, \gamma)/\gamma, \tag{3.8}$$

where $G$ is a Gaussian kernel to smooth the predicted mask, $*$ is the convolution operator, and $\gamma$ is a threshold designed to prevent assigning excessive weight to patches with significant deviations.

## 3.4 Experiments

### 3.4.1 Experimental setting

**Datasets.** We employ three datasets to asses the performance of UAD methods. **IXI Dataset** (IXI Dataset, 2004): a publicly available resource with brain MRI scans from approximately 600 healthy subjects. **ATLAS 2.0** (Liew *et al.*, 2022) includes 655 T1-weighted MRI scans accompanied by expert-segmented lesion masks. As a pre-processing, all brain scans of both IXI and ATLAS 2.0 datasets were registered to MNI152 1mm templates and normalized to the 98th

percentile. Then, mid-axial slices were extracted and padded to the resolution of 256×256 pixels. **BraTS'21** (Bakas *et al.*, 2017): following the experimental setup of IterMask$^2$ (Liang *et al.*, 2024b), we also employ this dataset, which comprises 1251 brain scans across four modalities: T1-weighted, contrast-enhanced T1-weighted (T1CE), T2-weighted, and T2 Fluid Attenuated Inversion Recovery (FLAIR). For each scan, 20 middle axial slices of the skull-stripped brain are extracted, which are padded to the resolution of 256×256 pixels.

**Training/Testing protocol.** We found that the training and testing protocols considerably differ in the UAD literature. For a fair comparison with prior methods, we evaluated our approach in two widely-adopted settings, comparing against the methods that were originally evaluated in each of these settings. **Setting-1 (*S1*)** (Bercea *et al.*, 2024a): training is performed on the middle slices of IXI subjects, whereas only middle slices of ATLAS 2.0 are used for testing. **Setting-2 (*S2*)** (Liang *et al.*, 2024b): in this setting, only normal slices from a given modality are used for training, while the abnormal slice of that modality with the largest pathology is employed for inference. The BRATS'21 dataset is used in this case, which is split into training (80%), validation (10%) and testing (10%).

**Evaluation metrics.** To evaluate the performance of our brain anomaly detection model, we use the Maximum Dice score, which reports the highest value obtained for thresholds ranging from 0 to 1. Following (Bercea *et al.*, 2024a), we employ the global Maximum Dice score in setting *S1*, which first flattens and concatenates all segmentations and predictions before calculating the maximum Dice score. For setting *S2*, we instead consider the regular Maximum Dice score.

**Implementation Details.** To project the data into the latent space, we employed a pre-trained perceptual compression VAE model (Rombach *et al.*, 2022). This model leverages an autoencoder trained using a combination of perceptual loss (Zhang *et al.*, 2018) and a patch-based adversarial objective, allowing for effectively reducing the spatial dimension by a factor of 8 ($256 \rightarrow 32$). As this model was originally trained for RGB images, we further adapted it and fine-tuned it for single-channel brain MRI data. Then, the VAE remained frozen throughout training the diffusion model (we used a UNet with attention as the diffusion model). The number of training

Table 3.1 **Performance in setting *S1:*** results across different lesion sizes, where bold highlights the best method and improvements of our approach compared to the best baseline are indicated in green

| Method | Pathology (Global Max Dice)↑ | | | |
|---|---|---|---|---|
| | Average | Small | Medium | Large |
| DDPM (Ho *et al.*, 2020b)$_{NeurIPS'20}$ | 8.1 | 1.4 | 9.5 | 25.7 |
| AnoDDPM (Wyatt *et al.*, 2022b)$_{CVPRw'22}$ | 18.1 | 4.8 | 23.5 | 46.7 |
| AutoDDPM (Bercea *et al.*, 2023a)$_{ICMLw'23}$ | 17.0 | 4.5 | 22.1 | 43.5 |
| pDDPM (Behrendt *et al.*, 2024b)$_{MIDL'24}$ | 22.3 | 8.0 | 30.2 | 47.7 |
| THOR (Bercea *et al.*, 2024a)$_{MICCAI'24}$ | 29.7 | 11.5 | 39.2 | 63.6 |
| **MAD-AD** *(Ours)* | **51.6**$_{+21.9}$ | **15.5**$_{+4.0}$ | **50.1**$_{+10.9}$ | **64.1**$_{+0.5}$ |

and inference time-steps ($T$) is set to 10. To form the random mask at each iteration, the masking ratio is drawn from a uniform distribution $U[0, 0.4]$, and the patch sizes of the mask along the X and Y axes are sampled independently from the following set: $\{1, 2, 4, 8\}$. The random mask is then multiplied by the brain mask to prevent noise in non-brain (i.e., background) patches. The model was trained for 300 epochs using a batch size of 96 and AdamW optimizer with a learning rate of $5 \times 10^{-4}$.

### 3.4.2 Results

**Quantitative results.** We empirically validated our method against a set of relevant state-of-the-art brain unsupervised anomaly detection methods in the two settings described in Section 3.4.1. Table 3.1 reports the results under the first setting, which uses middle slices of the IXI dataset for training, and middle slices of ATLAS 2.0 for evaluation. We can observe that the proposed approach substantially outperforms existing diffusion-based methods, particularly on small- and medium-sized lesions. More concretely, our approach improves the best baseline (the recent THOR method (Bercea *et al.*, 2024a)) by 4.0% and 10.9% in small and medium lesions, respectively, and by 21.9% when using the whole dataset (referred to as *"Average"*, as in (Bercea *et al.*, 2024a)). The performance gap further increases if we consider the second best baseline (i.e., pDDPM), where average differences are equal to nearly 30%. Note that even though our

Table 3.2 **Performance in setting *S2:*** results across different modalities, where bold highlights the best method and performance improvements (*resp.* decrease) of our approach compared to the best baseline are indicated in **green** (*resp.* **red**)

| Method | Modality (Max Dice)↑ | | | | |
|---|---|---|---|---|---|
| | FLAIR | T1CE | T2-w | T1-w | Avg |
| AE (Baur *et al.*, 2021a)$_{MedIA'21}$ | 33.4 | 32.3 | 30.2 | 28.5 | 31.1 |
| DDPM (Ho *et al.*, 2020b)$_{Neurips'20}$ | 60.7 | 37.9 | 36.4 | 29.4 | 41.1 |
| AutoDDPM (Bercea *et al.*, 2023a)$_{ICMLw'23}$ | 55.5 | 36.9 | 29.7 | 33.5 | 38.9 |
| Cycl.UNet (Liang *et al.*, 2023)$_{MICCAI'23}$ | 65.0 | 42.6 | 49.5 | 37.0 | 48.5 |
| DAE (Kascenas *et al.*, 2022)$_{[0,\infty]\ MIDL'22}$ | 79.7 | 36.7 | 69.6 | 29.5 | 53.9 |
| IterMask$^2$ (Liang *et al.*, 2024b)$_{MICCAI'24}$ | **80.2** | 61.7 | 71.2 | 58.5 | 67.9 |
| **MAD-AD** *(Ours)* | 76.2$_{\text{-4.0}}$ | **68.5**$_{\text{+6.8}}$ | **73.2**$_{\text{+2.0}}$ | **63.4**$_{\text{+4.9}}$ | **70.3**$_{\text{+2.4}}$ |

model yields superior performance for small pathologies, it still struggles to accurately locate these type of small abnormalities, similarly to existing approaches. In MAD-AD, this low performance may be due to the use of a diffusion model on a compressed latent space, which can lead to overlooking very small pathologies.

Under the second setting (*S2*), the proposed approach yields the best scores in three out of four modalities, leading to the highest average score (Table 3.2). While the differences with respect to the best baseline are smaller in this setting, improvements over the second best baseline are still considerably high, with an overall boost near to 16%. Thus, quantitative results under two common settings in the UAD literature demonstrate the superior performance of our approach for this task, highlighting its potential as a powerful alternative to existing methods.

**Ablation on using different sources for the anomaly score.** In this section, we investigate the impact of using different strategies to form the anomaly map: pixel-level discrepancies $(\boldsymbol{x}'_0, \boldsymbol{x}'_T)$, latent-space discrepancies $(\boldsymbol{z}'_0, \boldsymbol{z}'_T)$, and the average of the predicted mask at reverse diffusion steps $(\frac{1}{T}\sum_{t=1}^{T} f_{M,\theta}(\boldsymbol{z}'_t))$. These results, which are reported in Table 3.3, showcase the better performance of resorting to the image-level difference, motivating our design choice.

**Impact of hyper-parameters.** Next, we evaluate the influence of key hyper-parameters on the performance of the proposed method, whose results on the BRATS dataset are depicted in Table

Table 3.3 Effect of different sources for the anomaly score in MAD-AD (BRATS'21)

| Anomaly source | Modality (Max Dice)↑ | | | | |
|---|---|---|---|---|---|
| | T1-w | T1CE | T2-w | FLAIR | Avg |
| Average predicted mask | 60.4 | 62.3 | 65.6 | 66.1 | 63.6 |
| Latent-level diff | 63.0 | 66.2 | 69.6 | 75.5 | 68.6 |
| Image-level diff | **63.4** | **68.5** | **73.2** | **76.2** | **70.3** |

3.4. From these results, we can observe that the choices made for the hyper-parameters lead to the best results overall.

Table 3.4 Ablation study on two key hyper-parameters of MAD-AD

| Hyper-parameter | Value | Modality (Max Dice)↑ | | | | |
|---|---|---|---|---|---|---|
| | | T1-w | T1CE | T2-w | FLAIR | Avg |
| #DDIM steps | 2 | 62.3 | **70.1** | 68.5 | 75.3 | 69.0 |
| | 5 | 63.1 | 69.4 | 71.1 | 74.0 | 69.4 |
| | 10 | **63.4** | 68.5 | **73.2** | **76.2** | **70.3** |
| $\gamma$ | 0.2 | **63.4** | **68.5** | 73.2 | **76.2** | 70.3 |
| | 0.4 | 63.3 | 68.3 | **73.8** | 76.0 | **70.3** |
| | ✗ | 62.0 | 67.9 | 72.6 | 74.9 | 69.3 |

**Qualitative results.** To further highlight the effectiveness of our unsupervised anomaly detection method, we present qualitative results obtained on the ATLAS 2.0 dataset (*S1*) and across all modalities of the BraTS dataset (*S2*). Figure 3.3. Figure 3.3 showcases representative examples of anomalous instances, their normal counterpart reconstructions, segmentation, and anomaly map by MAD-AD. These qualitative results underscore the ability of our approach to accurately localize anomalous regions without relying on supervised labels.

## 3.5 Conclusion

This paper introduces a novel unsupervised anomaly detection method for brain MRI using a latent diffusion model with a random masking strategy. The approach leverages latent space, as brain anomalies in the latent space could be considered as noise and therefore be removed

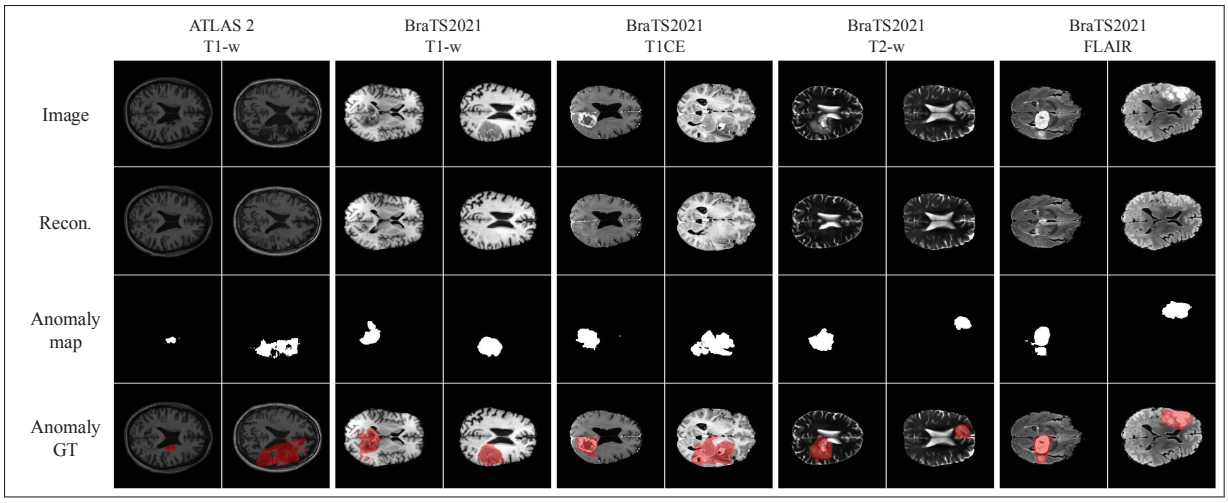



Figure 3.3 **Qualitative results.** Anomaly segmentation performance obtained by our approach (i.e., “Anomaly map") in brain MRI for different modalities and datasets

during the denoising process of diffusion models. Furthermore, by using a mask prediction module in the diffusion model, the model can selectively modify anomalous regions while preserving normal areas, enabling accurate identification of anomalous regions. Experiments on two datasets and two common brain UAD experimental settings demonstrate the superiority of our approach, validating its effectiveness in detecting and localizing brain anomalies without requiring labeled data, and showcasing its promising potential as an alternative to existing methods.

# CHAPTER 4

# CORRECTING DEVIATIONS FROM NORMALITY: A REFORMULATED DIFFUSION MODEL FOR MULTI-CLASS UNSUPERVISED ANOMALY DETECTION

Farzad Beizaee[1,2] , Gregory Lodygensky[2] , Christian Desrosiers[1] , José Dolz[1]

[1] ÉTS Montreal, Canada
[2] CHU Sainte-Justine, University of Montreal, Canada

Article accepted in the international conference of "Computer Vision and Pattern Recognition (CVPR)", February 2025

## 4.1 Introduction

Unsupervised anomaly detection (UAD) is paramount to a wide span of computer vision problems across strategic and high-impact domains, such as industrial inspection (Defard *et al.*, 2021; Liu *et al.*, 2023c), video surveillance (Pang *et al.*, 2020; Yan *et al.*, 2023), or medical imaging (Schlegl *et al.*, 2017; Silva-Rodríguez *et al.*, 2022). The main goal of UAD is to identify corrupted images, as well as anomalous pixels within these images, by leveraging only *normal* images. This setting naturally arises in many scenarios, where normal samples are readily available but compiling a curated set of labeled abnormal images is costly, due to the complexity of the annotation process as well as the scarcity and high variability of potential abnormalities.

The prevailing approach to address UAD frames this task as a *cold-start*[1] anomaly detection problem, which is typically approached through three primary categories of methods: reconstruction (Yan *et al.*, 2021; Perera, Nallapati & Xiang, 2019; Schlegl *et al.*, 2017), synthesizing (Li *et al.*, 2021; Zavrtanik, Kristan & Skočaj, 2021; Schlüter *et al.*, 2022; Zavrtanik, Kristan & Skočaj, 2022) and embedding (Defard *et al.*, 2021; Deng & Li, 2022; Roth *et al.*, 2022; Rudolph *et al.*, 2022) based approaches. While popular, single-class approaches hinder the scalability of these strategies, as the amount of storage and training time increases with the number of categories.

[1] Also commonly known as one-class classification (OCC).

Thus, there is a real need for novel methods that can accommodate the multi-class scenario in a robust and efficient manner.

Under the unified scenario, however, the distribution of normal data becomes more complex. As a result, the success of this task hinges on robust models capable of effectively learning the joint distribution across diverse types of objects. Diffusion models (Sohl-Dickstein *et al.*, 2015), and more particularly Denoising Diffusion Probabilistic Models (DDPM) (Ho *et al.*, 2020b), have emerged as strong candidates for this endeavor. Indeed, diffusion models have demonstrated strong potential in reconstructing the normal counterpart of an image to localize anomalous regions (Fučka, Zavrtanik & Skočaj, 2024; Zhang *et al.*, 2023a).

Diffusion models, though highly effective at generating high-quality samples, are designed to generate images from pure noise. Therefore, their direct application to modify the input images necessitates a forward diffusion process followed by a backward denoising process. However, this strategy imposes several limitations. First of all, directly using DDPMs in the multi-class context may result in misclassifying generated images due to the loss of their original category information (He *et al.*, 2024b), when a high number of diffusion steps is applied. On the other hand, if we use lower diffusion steps, the *"identity shortcut"* issue will emerge, where the model tends to simply denoise the input regardless of whether the content is normal or anomalous, thus preserving the anomalous regions(You *et al.*, 2022). A further problem arises when forward and reverse diffusion steps are applied indiscriminately across the entire image. In such cases, the model may struggle to fully recover the normal patterns leading to false identification of anomalies due to discrepancies between the input image and its reconstruction. This problem is more significant when the input image involves random patterns and textures, such as "tile". Lastly, the model's ability to accurately reconstruct the normal appearance of anomalous regions improves by leveraging the contextual information from surrounding normal regions. However, diffusion models introduce noise across the entire image, leading to the loss of crucial details required for reconstructing the normal counterpart of the image. These limitations are depicted in Figure 4.2.

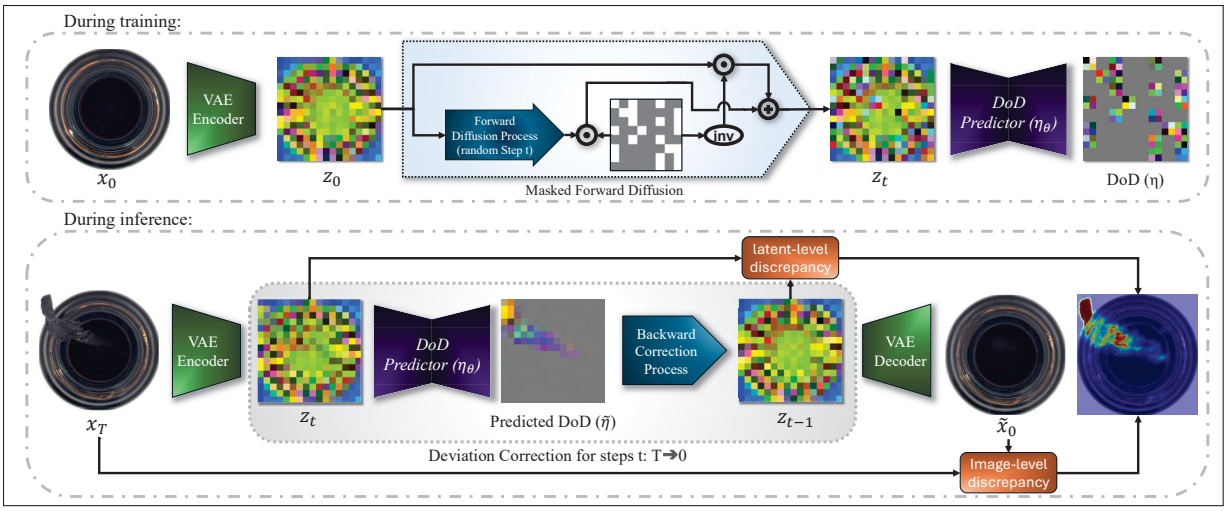



Figure 4.1 **Overview of the proposed method.** During **training** (*top*), normal images are partially diffused using random masks and randomly sampled time-steps ([1,$T$]). Then, our DeCo-Diff model is trained to predict the direction of deviation from the input image. At **inference** (*bottom*) starting from time-step $T$ for the target images, DeCo-Diff progressively corrects the deviation from normality

To address these limitations, we propose a reformulation of the standard diffusion models in such a way that it corrects the deviation from normality. This new formulation better suits our goal of reconstructing abnormal regions into their normal counterpart, while preserving the fine details of normal areas. Specifically, it enables us to directly use the target anomalous image in the backward correction process during inference, which eliminates the need for forward diffusion process, prevents the degradation of informative information, and enables selective correction of abnormal regions.The proposed model also introduces a random masking strategy for the added noise, which brings two important benefits. First, it leaves portions of the image untouched during the forward process, thus making the correction process conditioned on those normal untouched regions. Secondly, during the backward correction process, using our reformulation, the model learns to selectively alter abnormal regions of the image while keeping normal areas unchanged. This leads to a more precise anomaly detection and localization.

Our **key contributions** can be summarized as follows:

- We propose a novel reformulation of diffusion models that learns to correct deviation from the learned distribution of normality to its normal counterparts, rather than generating samples through denoising steps.
- We introduce a random masking strategy into the forward diffusion process, which conditions the deviation correction of abnormal regions to surrounding areas while preserving the fine details of normal regions.
- Furthermore, we reformulated the DDIM sampling (Song *et al.*, 2021a) to accommodate the deviation correction approach presented, enabling a faster, stable, and deterministic reverse process for efficient sampling.
- Extensive results on popular anomaly detection benchmarks demonstrate the superiority of our approach across recent state-of-the-art methods and multiple metrics.


Figure 4.2 **Diffusion model reconstruction vs. DeCo-Diff.** Fine details and patterns of a normal image are changed during the standard forward-backward diffusion process: the “wood" image becomes a “Tile" sample when $T$ steps are applied. In contrast, DeCo-Diff does not alter the details of the image using $T$ correction steps, maintaining the appearance of the original input image

## 4.2 Related work

### 4.2.1 Unsupervised Anomaly detection

Unsupervised anomaly detection has been studied from multiple perspectives: ***1) Reconstruction-based approaches*** rely on the assumption that a model trained on normal data will fail to reconstruct anomalous regions, leveraging the differences between the input and its reconstructed

image as an anomaly score. Generative Adversarial Networks (Yan *et al.*, 2021; Perera *et al.*, 2019; Schlegl *et al.*, 2017), Variational Auto-Encoders (Liu *et al.*, 2020), and Normalizing Flows (Gudovskiy *et al.*, 2022), are typically used as the backbone for reconstruction networks in this task. ***2) Synthesizing-based approaches*** typically introduce synthetic anomalies in normal images used for training (Li *et al.*, 2021; Zavrtanik *et al.*, 2021; Schlüter *et al.*, 2022; Zavrtanik *et al.*, 2022). For example, DRÆM (Zavrtanik *et al.*, 2021) trains a network on synthetically-generated just-out-of-distribution patterns, whereas CutPaste (Li *et al.*, 2021) introduces anomalies via a simple, yet efficient strategy based on cutting-pasting image patches at random locations on a normal image. ***3) Embedding-based approaches*** focus on embedding the normal features into a compressed space, relying on the assumption that anomalous features are far from the normal clusters in this space. These methods employ networks that are pre-trained on large-scale datasets like ImageNet for feature extraction (Defard *et al.*, 2021; Deng & Li, 2022; Roth *et al.*, 2022; Rudolph *et al.*, 2022). For instance, PaDiM (Defard *et al.*, 2021) resorts to a multivariate Gaussian to model the distribution of normal patch features at each position of the image, and then measures the normality score using the Mahalanobis distance. Similarly, PatchCore (Roth *et al.*, 2022) employs a core set to encode the features of normal patches, and finds anomalies at test time by computing the distance from a new patch's embedding and its nearest element in the core set. More recently, (Li *et al.*, 2024) proposed using the hyperbolic space to measure the distance between feature representations. Despite the progress made by these approaches in the single-class setting, their performance in the unified scenario remains unsatisfactory.

### 4.2.2 Multi-class unsupervised anomaly detection

The direct application of above-mentioned methods to the multi-class problem often leads to suboptimal performance. Moreover, since each category needs a separate trained model, the computational burden of these methods quickly explodes as the number of classes increases.

Tackling this more realistic scenario, multi-class unsupervised anomaly detection (Mc-UAD) approaches have recently gained traction within the community (You *et al.*, 2022; Deng & Li,

2022; He *et al.*, 2024b; Liu *et al.*, 2023c; Zhang *et al.*, 2023b; He *et al.*, 2024a; Zhang *et al.*, 2025; Guo *et al.*, 2023b). UniAD (You *et al.*, 2022) proposes a series of modifications to accommodate reconstruction-based networks for the challenging task of Mc-UAD. Observing the importance of the query embedding module in transformers to model the normal distribution, a query decoder is integrated into each layer, instead of only in the first layer as in vanilla transformers. To alleviate the "*identity shortcut*" problem in transformers, authors also introduce a neighbor mask attention module preventing tokens to copy themselves via self-attention. Last, a feature jittering strategy is employed to help the model recover normal features from noisy ones. Rd4AD (Deng & Li, 2022) leverages a reverse distillation approach, where a student network learns to restore the multi-scale representations of a teacher given the teacher one-class embeddings. MambaAD (He *et al.*, 2024a) proposed a pyramidal autencoder framework to reconstruct multi-scale features using recently proposed Vision Mamba networks (Zhu *et al.*, 2024). MoEAD (Meng *et al.*, 2024) introduced a Mixture-of-Experts architecture to transform single-class models into a unified model. DiAD (He *et al.*, 2024b) leveraged diffusion models for multi-class AD by integrating a semantic-guided network that helps preserve semantic information in the reconstruction of a Latent Diffusion Model. More Recently, GLAD (Yao *et al.*, 2024a) proposed combining the global and local adaptive mechanisms to improve the reconstruction performance of diffusion models. However, diffusion-based methods indiscriminately apply noise to the entire image and lack an explicit mechanism for learning how to reconstruct abnormal regions from their normal surrounding regions.

## 4.3 Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) are based on the idea of progressively perturbing data samples into noise via a forward process, and reversing this process to generate new data samples. This section introduces the fundamental concepts and notations required for understanding and applying diffusion models introduced in (Ho *et al.*, 2020b) and (Song *et al.*, 2021a).

### 4.3.1 Forward Diffusion Process

Let $\mathcal{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ denote the data samples (i.e., images), where $\boldsymbol{x}^{(i)} \in \mathbb{R}^{H \times W \times C}$ is the $i$-th image, with $H$, $W$, and $C$ representing its height, width and number of channels, respectively. The forward Markov diffusion process consists in gradually corrupting data samples by adding Gaussian noise for $t$ time-steps where $t \in [1, T]$. Following (Ho *et al.*, 2020b), the forward noising process can be characterized as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \;=\; \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_t, \beta_t\mathbf{I}). \tag{4.1}$$

Here, $\beta_i$ is a noise scheduling parameter that controls the amount of noise added at each step, such that $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for a sufficiently large $T$. The marginal distribution at the moment $t$ can then be explicitly defined from $\boldsymbol{x}_0$ as:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) \;=\; \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{4.2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{T} \alpha_i$.

### 4.3.2 Reverse Diffusion Process

In the reverse process, which iteratively removes noise through a series of denoising steps to generate new samples from Gaussian noise, the prior probability at each step can be modeled as a Gaussian distribution:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \;=\; \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \beta_t\mathbf{I}), \tag{4.3}$$

in which $\beta_t$ is fixed for each $t$, and the mean functions $\mu_\theta(\boldsymbol{x}_t, t)$ parameterized by $\theta$ are trainable:

$$\mu_\theta(\boldsymbol{x}_t, t) \;=\; \frac{1}{\sqrt{\bar{\alpha}_t}}\Big(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\Big). \tag{4.4}$$

This reverse process is trained with the variational lower bound of the log-likelihood of $\boldsymbol{x}_0$, i.e.:

$$\begin{aligned}\mathcal{L}(\theta) \;&=\; -\,p(\boldsymbol{x}_0|\boldsymbol{x}_1) \\ &\quad + \sum_t \mathcal{D}_{KL}\big(q^*(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\;\|\;p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)\big).\end{aligned} \tag{4.5}$$

By applying the reparameterization trick (i.e., parameterizing the noise function $\boldsymbol{\epsilon}_\theta$ as a neural network), the model can be trained using a simple mean-square error loss between the ground truth sampled Gaussian noise and the predicted noise:

$$\min_\theta \;\; \mathbb{E}_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t\right)\|_2^2\right], \tag{4.6}$$

with $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})$. Thus, starting from a noise vector $\boldsymbol{x}_T\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ sampling is achieved by iteratively applying the learned reverse process to gradually reconstruct a clean sample $\boldsymbol{x}_0$.

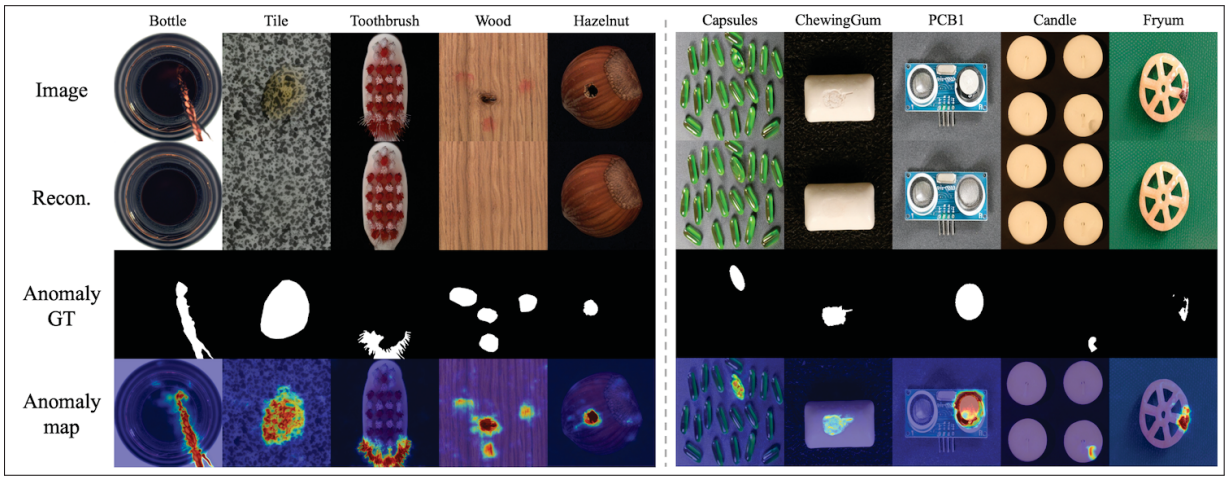



Figure 4.3 **Qualitative results.** From *top* to *bottom*: the original input image (with anomalies), DeCo-Diff reconstruction, the ground truth mask, and the predicted anomaly mask. Examples are depicted for two datasets (MVTec-AD on the left side and VisA on the right side) and across multiple anomalies with diverse complexity

## 4.4 Method

We resort to diffusion models to characterize the distribution of normal images, and reconstruct the normal counterparts of target images to detect and localize anomalies. However, standard diffusion models are inherently designed to generate new samples starting from pure noise and not altering selective regions of an input image. Moreover, by applying the forward and then reverse diffusion processes to reconstruct the normal counterpart of an anomalous image, depending on time-step $t$, the information in the normal regions can be partially degraded or lost. This degradation leads to a suboptimal reconstruction of normal areas, which is not ideal for unsupervised anomaly detection as input images and their reconstructed counterparts are compared to detect abnormal regions.

To address this issue, we propose a modified diffusion model (DeCo-Diff) that selectively alters only the anomalous regions of an image, while leaving the normal areas intact. This approach allows for preserving the normal parts of anomalous images while altering abnormal regions, based on the learned distribution of normal images. In this section, we detail the different components of our approach, whose overall pipeline is depicted in fig. 4.1.

### 4.4.1 Modeling anomalies as noise in latent space

We employ a pre-trained Variational Auto-Encoder (VAE) (Rombach *et al.*, 2022) to project images into a latent space where the diffusion process is performed. We chose to work in this latent space for five key reasons: *i)* anomalies in the image can be effectively interpreted as noise within the latent space, which aligns well with the operational framework of diffusion models; it also *ii)* alleviates the problem of "identity shortcut" which causes anomalies to be preserved in the reconstruction; *iii)* increases computational efficiency; *iv)* enhances stability, especially when training data is limited; and *v)* improves the quality of generated samples.

Denoting the encoder and decoder networks of the VAE as $\phi_E$ and $\phi_D$, an input image $\boldsymbol{x}^{(i)}$ is projected into a latent space representation $\boldsymbol{z}^{(i)} = \phi_E(\boldsymbol{x}^{(i)})$, where $\boldsymbol{z}^{(i)} \in \mathbb{R}^{H' \times W' \times C'}$. It is

important to note that, due to the unsupervised nature of this method, we only have access to normal images during the training phase.

*Limitations of standard diffusion models.* There are certain limitations to adding noise to the whole input image (or its latent representation) as in standard diffusion models. First, the added noise also affects normal regions. Consequently, during the forward diffusion process, normal regions may experience a partial or complete loss of information, making it difficult to fully recover the input image and potentially leading to their misinterpretation as anomalous areas when compared to the input image. Furthermore, abnormal regions should be reconstructed with respect to the surrounding normal areas. However, if normal patches are altered due to the forward diffusion process, they cannot be fully exploited to reconstruct abnormal patches, a critical step for anomaly detection.

To alleviate the aforementioned limitations, we propose to integrate random masking into the forward diffusion process. Given the latent features of an input normal sample $z_0 \sim p\,(z_0)$, we spatially divide $z_0$ into non-overlapping noisy $z_0^n$ (non-masked) and visible $z_0^v$ patches (masked) using a random mask with a random masking ratio ($r_{mask}$). Afterward, during the forward process, only the noisy patches in $z_0^n$ are gradually diffused, while the visible patches $z_0^v$ remain unchanged:

$$q(z_t^k|z_0^k) \;=\; \begin{cases} \mathcal{N}(z_t^n; \sqrt{\bar{\alpha}_t} z_0^n, (1-\bar{\alpha}_t)\mathbf{I}), & \textit{if } k = n \\ z_0^v, & \textit{otherwise.} \end{cases} \tag{4.7}$$

This way, visible patches in the image can represent normal regions that have been unaltered during the forward process, while noisy patches could mimic anomalous areas.

### 4.4.2 Deviation instead of Diffusion

While we motivated the use of a masking strategy (eq. (4.7)) during the forward diffusion process, conditioning the reverse process and training objective on the masks used for the forward pass is suboptimal and impractical. First, accessing to prior knowledge on the location of abnormal regions during test time is unrealistic. Secondly, the denoising operation in eq. (4.3) is not

applicable to *non-noisy* patches. Thus, we need to integrate a mechanism that optimizes the training objective and applies the reverse process to both noisy and visible patches.

To achieve this, we reformulate the standard diffusion model to better suit our needs. Let us first consider the noisy patches. Based on eq. (4.1), for noisy patches in the latent space we have:

$$z_t^n = \sqrt{1-\beta_t} z_{t-1}^n + \sqrt{\beta_t}\boldsymbol{\epsilon} \tag{4.8}$$

Based on eq. (4.2), $z_t^n$ could be explicitly obtained with

$$z_t^n = \sqrt{\bar{\alpha}_t} z_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{4.9}$$

We can then rewrite eq. (4.9) as:

$$\begin{aligned} z_t^n &= (1-(1-\sqrt{\bar{\alpha}_t}))z_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \\ &= z_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} - (1-\sqrt{\bar{\alpha}_t})z_0^n \\ &= z_0^n + \underbrace{\sqrt{1-\bar{\alpha}_t}\left(\boldsymbol{\epsilon} - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} z_0^n\right)}_{\textit{“Deviation from Normality”}} \end{aligned} \tag{4.10}$$

In the above equation, the second term measures the deviation of $z_t^n$ from $z_0^n$, which represents a normal image. In summary, we have that:

$$\begin{aligned} z_t^n &= z_0^n + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta} \\ \boldsymbol{\eta} &= \left(\boldsymbol{\epsilon} - \frac{1-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}} z_0^n\right), \end{aligned} \tag{4.11}$$

where $\boldsymbol{\eta}$ is a term based on the time-step $t$, $z_0$, and $\boldsymbol{\epsilon}$, which points to the “*Direction of Deviation*” (DoD). One interesting characteristic of this approach is that the patches will remain untouched in the forward process if $\boldsymbol{\eta} = \mathbf{0}$.

The reverse process aims to correct the deviation from a normal image by progressively removing noise through a learned denoising function $p(z_{t-1}|z_t)$. This could be achieved by training a neural network to closely approximate the true reverse path by predicting the DoD at each time-step. The main objective can thus be defined as:

$$\min_{\theta} \ \mathbb{E}_{z_0 \sim q(z_0), \epsilon, t} \left[ \| \boldsymbol{\eta} - \boldsymbol{\eta}_\theta \left( z_t, t \right) \|_2^2 \right]. \tag{4.12}$$

This approach encourages the model to predict the DoD from normality at each time-step, facilitating accurate correction in the reverse process. As a result, our DoD predictor ($\boldsymbol{\eta}_\theta$) learns a robust denoising trajectory that can correct the noisy, abnormal patches with respect to the surrounding normal ones, while keeping the normal patches untouched by simply predicting zero as their DoD. Furthermore, to expose the model to more structured anomalies during training, we also propose to incorporate a strategy based on *patch shuffling*. Since the latent space follows a normal distribution, for a portion of patches (specified with $r_{shuffle}$), instead of adding Gaussian noise, we can replace them with other patches taken from images in the same batch. This strategy could ultimately result in better localization of large and structured anomalies.

Table 4.1 **Quantitatve evaluation on MVTec-AD.** Image and Pixel-level results on *multi-class* anomaly detection. Best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best approach. Differences with the second (or best) approach in (gray)

| Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | $f1_{max}$ | AUROC | AUPRC | $f1_{max}$ | AUPRO |
| RD4AD *CVPR'22* | 94.6 | 96.5 | 95.2 | 96.1 | 48.6 | 53.8 | 91.1 |
| UniAD *NeurIPS'22* | 96.5 | 98.8 | 96.2 | 96.8 | 43.4 | 49.5 | 90.7 |
| SimpleNet *CVPR'23* | 95.3 | 98.4 | 95.8 | 96.9 | 45.9 | 49.7 | 86.5 |
| DeSTSeg *CVPR'23* | 89.2 | 95.5 | 91.6 | 93.1 | 54.3 | 50.9 | 64.8 |
| DiAD *AAAI'24* | 97.2 | 99.0 | 96.5 | 96.8 | 52.6 | 55.5 | 90.7 |
| MoEAD *ECCV'24* | 98.0 | 99.3 | 97.5 | 96.9 | 49.8 | 43.5 | 91.0 |
| GLAD *ECCV'24* | 97.5 | 99.1 | 96.6 | 97.4 | **60.8** | **60.7** | 93.0 |
| MambaAD *NeurIPS'24* | **98.6** | **99.6** | **97.8** | **97.7** | 56.3 | 59.2 | **93.1** |
| DeCo-Diff (*Ours*) | **99.3**$_{+0.7}$ | **99.8**$_{+0.2}$ | **98.5**$_{+0.7}$ | **98.4**$_{+0.7}$ | **74.9**$_{+14.1}$ | **69.7**$_{+9.0}$ | **94.9**$_{+1.8}$ |

### 4.4.3 Anomaly detection

In order to detect anomalies at inference time, we first need to adapt the sampling of DDPM (eq. (4.3)) to fit the proposed configuration for correcting the deviation from normality. However, before describing this step, it is worth mentioning that sampling with DDPMs requires many reverse sampling steps until $t = 0$, leading to a significant computational burden. In addition, DDPMs introduce noise over the whole image at each step $t$, which goes against our objective of keeping normal patches unchanged. Consequently, we resort to the Denoising Diffusion Implicit Model (DDIM) (Song *et al.*, 2021a) during inference, which modifies the sampling process by making it deterministic instead of stochastic. To correct the deviation using DDIM sampling, with a trained model $\boldsymbol{\eta}_\theta$, the predicted normal image ($\tilde{z}_0$) at each time-step $t$ is:

$$\tilde{z}_0 \;=\; z_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\eta}_\theta(z_t, t). \tag{4.13}$$

Therefore, the reverse process becomes:

$$\tilde{z}_{t-1} \;=\; \tilde{z}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \;\; \underbrace{\boldsymbol{\eta}_\theta(z_t, t)}_{\text{“}\textit{Direction of deviation}\text{"}} \tag{4.14}$$

In the above equation, $\tilde{z}_0$ is the predicted $z_0$, and $\boldsymbol{\eta}_\theta(z_t, t)$ the DoD. Note that eq. (4.14) enables a deterministic trajectory from $z_\tau$ to $z_0$, which accommodates our setting since it avoids introducing any noise. It is noteworthy to mention that a key advantage of DDIM over DDPM is that, by eliminating the noise at each step and making the denoising process more direct, high-quality samples can be generated in fewer steps. This results in a faster convergence to normal samples.

Intuitively, as our DeCo-Diff model has been solely trained using normal images, when an anomalous image is provided the anomalous regions will fall outside the learned distribution. Consequently, the model will consider them as non-masked noisy patches, which will be denoised during the reverse process. On the other hand, if the patches do not correspond to anomalous regions, the model will consider them as masked normal patches and they will remain unchanged

during the reconstruction. Furthermore, in contrast to standard diffusion processes that add noise to the whole image, the proposed model allows leveraging surrounding normal information when reconstructing abnormal regions.

**Leveraging multi-scale information.** Pixel-level reconstruction discrepancy is widely used in reconstruction-based UAD approaches mainly due to this ability to produce detailed anomaly maps. Nevertheless, some anomalous regions might not be revealed using this approach, for example, if they only present small differences in color compared to the normal ones. On the other hand, reconstruction discrepancy measured in the latent space can capture more subtle anomalies, however, it yields coarser anomaly maps due to the reduced spatial dimension of the latent space.

Based on these observations, we propose a multi-scale strategy where discrepancies at both pixel-space and latent-space are leveraged jointly to localize potential anomalies. More concretely, in addition to the reconstructed "normal" feature embedding $\tilde{\boldsymbol{z}}_0$, we also generate a reconstructed normal sample in the image-space as $\tilde{\boldsymbol{x}}_0 = \phi_D(\tilde{\boldsymbol{z}}_0)$, where $\phi_D$ is the pre-trained VAE decoder. The predicted anomaly map can be thus obtained using the geometric mean of latent-level and pixel-level discrepancies as:

$$\boldsymbol{a} = \sqrt{\frac{1}{\gamma_l \cdot \gamma_p} \min(\|\tilde{\boldsymbol{z}}_0 - \boldsymbol{z}_0\|, \gamma_l) \cdot \min(\|\tilde{\boldsymbol{x}}_0 - \boldsymbol{x}_0\|, \gamma_p)}. \tag{4.15}$$

Here, $\gamma_l$ and $\gamma_p$ respectively denote the latent-level and pixel-level thresholds, which prevent assigning excessive weight to patches with significant deviations in either the image or the latent space. For instance, if a black patch is reconstructed as white, this does not necessarily indicate that the patch is more anomalous compared to a patch reconstructed as red.

Table 4.2 **Quantitatve evaluation on VisA.** Image and Pixel-level results on *multi-class* anomaly detection. The best method (per metric) is highlighted in **blue**, whereas **red** is used to denote the second-best approach. Differences with second (or best) approach in (gray)

| Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|
| | AUROC | AUPRC | $F1_{max}$ | AUROC | AUPRC | $F1_{max}$ | AUPRO |
| RD4AD *CVPR'22* | 92.4 | 92.4 | 89.6 | 98.1 | 38.0 | 42.6 | 91.8 |
| UniAD *NeurIPS'22* | 88.8 | 90.8 | 85.8 | 98.3 | 33.7 | 39.0 | 85.5 |
| SimpleNet *CVPR'23* | 87.2 | 87.0 | 81.8 | 96.8 | 34.7 | 37.8 | 81.4 |
| DeSTSeg *CVPR'23* | 88.9 | 89.0 | 85.2 | 96.1 | **39.6** | 43.4 | 67.4 |
| DiAD *AAAI'24* | 86.8 | 88.3 | 85.1 | 96.0 | 26.1 | 33.0 | 75.2 |
| MoEAD *ECCV'24* | 93.0 | **95.1** | **89.8** | **98.7** | 36.6 | 41.0 | 88.6 |
| GLAD *ECCV'24* | 91.8 | 92.9 | 88.0 | 97.4 | 34.6 | 40.3 | **92.0** |
| MambaAD *NeurIPS'24* | **94.3** | 94.5 | 89.4 | 98.5 | 39.4 | **44.0** | 91.0 |
| DeCo-Diff (*Ours*) | **96.4**$_{+2.1}$ | **96.8**$_{+1.7}$ | **92.2**$_{2.4}$ | **98.5**$_{-0.2}$ | **51.3**$_{+11.7}$ | **51.2**$_{+7.2}$ | **92.1**$_{+0.1}$ |

## 4.5 Experiments

### 4.5.1 Setting

**Datasets.** We evaluate our method on two well-known anomaly detection datasets. **MVTec-AD (Bergmann *et al.*, 2019)** simulates real-world industrial production scenarios, filling the gap in unsupervised anomaly detection. It consists of 5 types of textures and 10 types of objects, in 5,354 high-resolution images from different domains. **VisA (Zou *et al.*, 2022)** consists of 10,821 high-resolution images, with 78 types of anomalies. It comprises 12 subsets corresponding to distinct objects categorized into three object types: Complex structure, Multiple instances, and Single instance. Note that ablations are performed on MVTec-AD.

**Evaluation Metrics.** Following prior literature (He *et al.*, 2024b,a), we use Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under Precision-Recall Curve (AUPRC[2]), and F1-score-max ($F1_{max}$) to measure performance in both anomaly detection (i.e., *image-level*) and anomaly localization (i.e., *pixel-level*). Furthermore, Area Under Per-Region-Overlap (AUPRO) is employed for pixel-level anomaly localization.

[2] Also called Average Precision(AP) or AUPR in the literature

**Baselines.** We compare our approach to eight recently-proposed approaches for multi-class UAD: RD4AD (Deng & Li, 2022), UniAD (You *et al.*, 2022), SimpleNet (Liu *et al.*, 2023c), DeSTSeg (Zhang *et al.*, 2023b), DiAD (He *et al.*, 2024b), MoEAD(Meng *et al.*, 2024), GLAD (Yao *et al.*, 2024a), and MambaAD (He *et al.*, 2024a).

**Implementation Details.** To map data into the latent space using a pre-trained VAE, we used a pre-trained perceptual compression model (Rombach *et al.*, 2022), which consists of an auto-encoder trained by a combination of a perceptual loss (Zhang *et al.*, 2018) and a patch-based adversarial objective that down-samples the image by a factor 8. Similar to VAEs(Kingma, 2013), a regularization loss measuring the KL divergence between latent features and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ is used to avoid the latent space having an arbitrarily high variance.

For our model $\boldsymbol{\eta}_\theta$ predicting the *direction of deviation* (DoD), following (Rombach *et al.*, 2022), we employed an attention UNet architecture with timestep embedding and a squared-cosine-beta scheduler (Nichol & Dhariwal, 2021). At each iteration and for each training sample, we first randomly select a timestep $t$ from a uniform distribution between 1 and $T$ (where $T$=10). Then, we sample a masking ratio ($r_{\text{mask}}$) from a uniform distribution $U[0, 0.7]$ and a latent patch size from $\{1, 2, 4, 8\}$ corresponding to pixel-wise patch sizes of {8×8, 16×16, 32×32, 64×64}. Using these sampled values, we generate a random mask with the specified ratio. Moreover, in our implementation, the replacement ratio ($r_{shuffle}$) is sampled in each iteration from $U[0, 0.3]$. Finally, $z_t$ is obtained using eq. (4.7), and serves as input to train our model. This model was trained 800 epochs for experiments on the MVTec dataset and 200 epochs for VisA, in both cases using a batch size of 128 and a single A6000 GPU. We used the AdamW optimizer with a Cosine Annealing scheduler with warm-up, with an initial learning rate set to $10^{-4}$, decaying to a minimum of $10^{-5}$. The VAE remained frozen throughout training.

### 4.5.2 Results

#### 4.5.2.1 Quantitative Results

***MVTec-AD:*** table 4.1 reports the performance for all methods, where we can see that the proposed DeCo-Diff obtains the overall best performance across all metrics for both image-wise and pixel-wise multi-class anomaly detection. More concretely, DeCo-Diff achieves 0.7%, 0.2%, 0.7% improvements over the second best approach, i.e., MambaAD (He *et al.*, 2024a), for image-level AUROC, AUPRC, and $F1_{max}$, respectively. The boost in performance is even more notable at pixel-level, an arguably more complex scenario. In particular, our method yields improvements of 0.2%, 14.1%, 9.0%, and 1.8% over the existing state-of-the-art multi-class UAD models measured by pixel-level AUROC, AUPRC, $F1_{max}$, and AUPRO, respectively. ***VisA:*** Results in table 4.2 confirm the trend observed in MVTec-AD, with our approach outperforming recent methods in both pixel-level and image-level metrics. Particularly, DeCo-Diff obtains around 2% performance gain over the second-best approach in image-level metrics, and exhibits significant pixel-level improvement over all methods in terms of AUPRC, $F1_{max}$, and AUPRO, yet ranking second for pixel-level AUROC.

#### 4.5.2.2 Qualitative Results

The quantitative results presented above are supported by visual examples showcasing the effectiveness of our approach in identifying anomalous image regions. In fig. 4.3, we depict anomalous samples from both datasets and their reconstruction using our model. As it can be observed, abnormal regions are properly modified during the deviation correction process and replaced by their normal counterpart. In contrast, the normal regions remained unchanged, our model preserving the fine details in these regions. This enables a precise detection and localization of abnormalities when contrasting the input image to its reconstructed version (fig. 4.3, *last row*). Moreover, fig. 4.4 visually depicts the deviation correction reverse process from DeCo-Diff, which showcases how abnormal areas are progressively removed, until a *cleaned* version of the input image is generated.

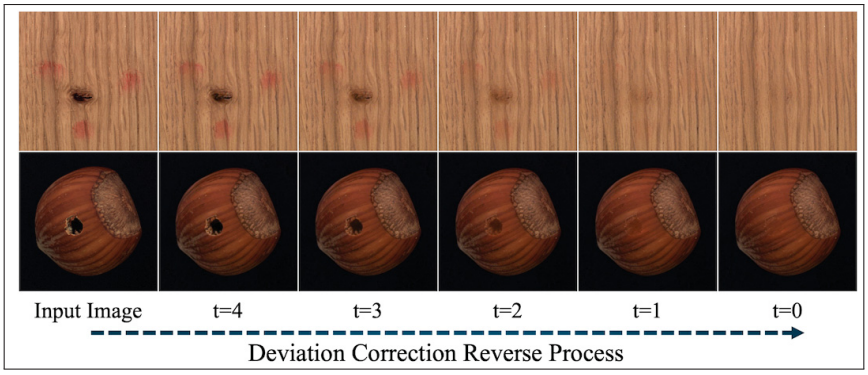



Figure 4.4 **Visualization of deviation correction through the reverse process steps.** Abnormal areas are progressively corrected, while fine normal details are preserved during the reverse process

#### 4.5.2.3 Ablation Studies

**Effect of backward deviation correction strategy and number of time-steps.** In this section, we explore different strategies for the deviation correction process during inference. In eq. (4.14), we explained how the proposed deviation correction formulation can be used in the DDIM reverse sampling to progressively correct the abnormalities in each time-step $t$. An alternative strategy would be to directly replace $z_{t-1}$ with $\tilde{z}_0$ at each time-step. This approach allows the model to correct the input image based on its initial prediction at the first time-step, and then continue refining any remaining abnormalities in subsequent steps.

We evaluated the performance for different number of reverse steps using both approaches, where the first approach is referred to as $\tilde{z}_{t-1}$ and the latter as $\tilde{z}_0$ in table 4.3. Note that, for a single reverse step, both approaches are identical and yield the same results. As reported in table 4.3, the first strategy ($\tilde{z}_{t-1}$) performs better in image-level metrics, while the second yields slightly better performance in terms of pixel-level metrics. Furthermore, the performance is generally consistent as the number of reverse time-steps increases, making a lower number of time-steps (but larger than 1) more favorable due to its decreased computational cost.

Table 4.3 Impact of the correction strategy at each time-step and number of reverse steps

| Correction strategy | Reverse steps | Image-level AUROC / AUPRC / $F1_{mask}$ | Pixel-level AUROC / AUPRC / $F1_{mask}$ / AUPRO |
|---|---|---|---|
| $\tilde{z}_0$ | 1 | 98.4 / 99.3 / 97.9 | 98.4 / 72.9 / 68.8 / 94.3 |
| $\tilde{z}_0$ | 2 | 99.2 / 99.7 / 98.1 | **98.4** / **75.1** / **69.9** / 94.7 |
| $\tilde{z}_0$ | 5 | 99.2 / 99.7 / 97.9 | 98.2 / 74.3 / 69.3 / 94.6 |
| $\tilde{z}_0$ | 10 | 98.9 / 99.5 / 97.6 | 97.8 / 73.6 / 68.8 / 94.2 |
| $\tilde{z}_{t-1}$ | 2 | 99.3 / 99.8 / 98.3 | 98.4 / 75.0 / 69.8 / 94.4 |
| $\tilde{z}_{t-1}$ | 5 | **99.3** / **99.8** / **98.5** | 98.4 / 74.9 / 69.7 / **94.9** |
| $\tilde{z}_{t-1}$ | 10 | 99.3 / 99.8 / 98.5 | 98.4 / 74.9 / 69.8 / 95.0 |

**Effect of leveraging discrepancies at pixel, latent, and visual feature levels.** In reconstruction-based anomaly detection, mapping the differences between inputs and their reconstructions to anomaly maps is essential. Thus, we now investigate several alternatives to the proposed strategy, presented in eq. (4.15). First, we consider using only the difference in pixel values, i.e., $\Delta \boldsymbol{x} = |\boldsymbol{x}_T - \boldsymbol{x}_0|$, which are clipped and scaled by $\gamma_l = 0.4$, as: $\frac{1}{\gamma_l} \min(\Delta \boldsymbol{x}, \gamma_l)$. However, this option may not be optimal, as anomalies might appear with the same color as normal regions. Then, we resort to the difference in the latent space ($\Delta \boldsymbol{z} = |\boldsymbol{z}_T - \boldsymbol{z}_0|$), which is similarly clipped and scaled by the same value, and then resized to match the target image size. Additionally, we explored the impact of the arithmetic mean: $\frac{1}{2\gamma_l} \min(\Delta \boldsymbol{z}, \gamma_l) + \frac{1}{2\gamma_p} \min(\Delta \boldsymbol{x}, \gamma_p)$. Finally, we extracted the visual features from a pre-trained ResNet50 (He *et al.*, 2016) for both the input images and their reconstructions and compared them using the cosine similarity to localize anomalies. The results for anomaly detection and localization using different levels of discrepancies are presented in table 4.4. These results support our approach (i.e., Geometric mean, eq. (4.15)), and demonstrate that combining pixel- and latent-space discrepancies with the geometric mean yields superior performance to considering these discrepancies separately.

## 4.6 Conclusions

We proposed a reformulation of standard diffusion models, specifically designed to modify abnormal regions in target images without affecting normal areas. By integrating masked strategy

Table 4.4 Different levels of discrepancy. Results on MVTec-AD

| Dissimilarity | Image-level AUROC / AUPRC / $F1_{mask}$ | Pixel-level AUROC / AUPRC / $F1_{mask}$ / AUPRO |
|---|---|---|
| Pixel diff. ($\Delta \boldsymbol{x}$) | 98.9 / 99.6 / 97.7 | 97.5 / 68.0 / 63.7 / 92.1 |
| Latent diff. ($\Delta z$) | 99.2 / 99.7 / 98.3 | 98.2 / 70.5 / 67.5 / 93.5 |
| Arithmetic ($\Delta \boldsymbol{x}$, $\Delta z$) | **99.3** / **99.8** / **98.5** | 98.2 / 73.5 / 68.8 / 94.1 |
| Geometric ($\Delta \boldsymbol{x}$, $\Delta z$) | **99.3** / **99.8** / **98.5** | **98.4** / **74.9** / **69.7** / **94.9** |
| Features (ResNet50) | 98.5 / 99.3 / 97.3 | 97.4 / 66.7 / 64.6 / 91.7 |

to the new formulation in the latent space, DeCo-Diff progressively detects and corrects the deviations from normality, enabling precise localization of abnormalities. Our comprehensive Quantitative and Qualitative results demonstrate the superiority of our model compared to state-of-art UAD methods in the unified multi-class context.

# CHAPTER 5

# REFLECT: RECTIFIED FLOWS FOR EFFICIENT BRAIN ANOMALY CORRECTION TRANSPORT

Farzad Beizaee[1,2] , Sina Hajimiri[1] , Ismail Ben Ayed[1] , Gregory Lodygensky[2] , Christian Desrosiers[1] , José Dolz[1]

[1] ÉTS Montreal, Canada
[2] CHU Sainte-Justine, University of Montreal, Montreal, Canada

Article accepted in the international conference of "Medical Image Computing and Computer Assisted Intervention (MICCAI)", June 2025

## 5.1 Introduction

Brain anomaly detection using medical images is a critical task in neuroimaging, with significant implications for early diagnosis and treatment planning. Brain abnormalities such as tumors, lesions, or traumatic injuries often appear as structural deviations from normal anatomy. While supervised methods for localizing anomalies are effective in many settings, they rely on large annotated datasets, which are costly and often scarce, particularly for rare anomalies. This has led to growing interest in unsupervised methods, where models learn the normal data distribution and subsequently localize deviations as potential abnormalities.

In recent years, generative models have been widely investigated for reconstruction-based unsupervised anomaly detection (UAD). Initial attempts leverage Auto-Encoders (AEs) (Baur *et al.*, 2021a; Baur *et al.*, 2019) and their variants, including Variational Auto-Encoders (VAEs) (Silva-Rodríguez *et al.*, 2022; Zimmerer *et al.*, 2019). Generative Adversarial Networks (GANs) (Goodfellow *et al.*, 2014), including AnoGAN (Schlegl *et al.*, 2017) and f-AnoGAN (Schlegl *et al.*, 2019a), have also emerged as promising alternatives to AE-based methods. However, these approaches tend to overfit the normal training data or yield blurry reconstructions. Normalizing flows (NFs) (Tabak & Turner, 2013; Rezende & Mohamed, 2015) are another family of generative models that transform a simple base distribution into a complex target distribution through a series of invertible transformations, thereby allowing exact

likelihood estimation. This makes them particularly appealing for anomaly localization via out-of-distribution detection Gudovskiy *et al.* (2022); Kim, Baik & Kim (2023); Chiu & Lai (2023); Zhao, Ding & Zhang (2023). However, NFs require complex architectures, are computationally expensive, and often involve iterative steps.

Progress in generative modeling has led to the rise of diffusion models (Ho *et al.*, 2020b), a powerful class of probabilistic models that generate high-quality data by gradually transforming noise into structured outputs through a learned iterative process. Fueled by their impressive generative performance, diffusion models have been increasingly adopted in medical imaging tasks, including unsupervised anomaly detection (Behrendt *et al.*, 2024a; Behrendt *et al.*, 2024c; Beizaee *et al.*, 2025; Bercea *et al.*, 2024a; Liang *et al.*, 2024a; Marimont *et al.*, 2024; Naval Marimont *et al.*, 2024; Wyatt *et al.*, 2022b). AnoDDPM (Wyatt *et al.*, 2022b) resorts to a partial diffusion strategy, where it adds noise to the image up to a certain timestep and then recovers it via reverse diffusion, whereas pDDPM (Behrendt *et al.*, 2024a) applies diffusion patch-wise to better capture local context. THOR (Bercea *et al.*, 2024a) refines diffusion models by using implicit temporal guidance via anomaly maps during the reconstruction process. Inspired by diffusion models, Itermask (Liang *et al.*, 2024a) proposes iterative mask refinement using reconstruction errors to better localize brain lesions in MRI. And very recently, MAD-AD (Beizaee *et al.*, 2025) treats abnormalities as noise in the latent space and uses a masked diffusion process to selectively correct abnormalities. While diffusion models offer a better overall performance, they tend to "memorize" patterns from training data (Somepalli *et al.*, 2023), which reduces their generalizability. Also, they need many iterative steps to reconstruct normal images, even with DDIM sampling (Song *et al.*, 2021a). Last but not least, all these generative methods are primarily designed to generate new samples (often from pure noise), and not to modify existing images. This limits their applicability for the selective correction of images.

An alternative unsupervised approach uses self-supervised learning to restore normal images corrupted with synthetic anomalies. Methods like Foreign Patch Interpolation (FPI)(Tan *et al.*, 2022) and Poisson Image Interpolation (PII)(Tan *et al.*, 2021) introduce such defects by blending

patches from different normal images, enabling pixel-level anomaly localization. Most recently, DISYRE v2 (Naval Marimont *et al.*, 2024) introduced a cold-diffusion pipeline that restores synthetically corrupted images with controlled anomaly severity through iterative refinement. While these methods are effective, their reliance on synthetic anomalies limits generalization, as these may not fully capture the variability of real-world abnormalities.

To address the aforementioned limitations, we propose REFLECT, an unsupervised brain anomaly detection framework built on the recently introduced rectified flows (Liu *et al.*, 2023a). Rectified flows, which learn a transport map between two different distributions through rectified trajectories, exhibit several advantages over the previously mentioned generative models, such as improved stability, high-fidelity reconstructions, and direct and efficient mapping, which enables the mapping of a sample to a target distribution with a single step.

Our work makes the following key contributions: *i*) We propose leveraging rectified flows in latent space to enable optimal and straight transport of abnormal brain samples toward their normal counterparts, thereby facilitating unsupervised brain anomaly detection with enhanced accuracy and reliability. To the best of our knowledge, rectified flow has not been explored before for unsupervised anomaly detection. By learning straight flow trajectories, our approach requires fewer time steps, which enables high-quality correction of abnormal regions in a single step while preserving normal regions. Unlike the often complex formulations used in diffusion models or GANs, rectified flows' objective directly controls the geometry of the flow and provides a natural mechanism to correct anomalous samples, making the method both theoretically elegant and practically efficient. *ii*) Moreover, we introduce an effective technique for generating diverse and realistic anomalous brain MRIs using a random walk-based masking strategy operating in the latent space. Masked regions are replaced with textured image segments or random noise, enhancing the variability and realism of synthesized anomalies. *iii*) Extensive experiments on brain anomaly detection benchmarks further demonstrate the superiority of our approach over recent state-of-the-art unsupervised anomaly detection methods.

## 5.2 Preliminaries: Rectified Flows

Rectified flows (Liu *et al.*, 2023a), building upon the principles of continuous normalizing flows (Chen *et al.*, 2018), learn an ordinary differential equation (ODE) that transports samples from an initial distribution $\pi_0$ to a target distribution $\pi_1$ along trajectories that are as straight as possible in a continuous-time framework. Concretely, consider the linear interpolation between random variables $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$:

$$X_t = (1 - t)\, X_0 + t\, X_1, \quad t \in [0, 1]. \tag{5.1}$$

Although $X_t$ provides a continuous trajectory from $X_0$ to $X_1$, these straight-line paths are non-causal and may intersect when considering different sample pairs. Such intersections are undesirable for generative modeling, as they lead to non-causal and non-deterministic dynamics. This interpolation is "rectified" into a causal ODE flow $\{Z_t : t \in [0, 1]\}$ via

$$\frac{dZ_t}{dt} \;=\; v(Z_t, t), \quad Z_0 \sim \pi_0, \tag{5.2}$$

where $Z_t$ represents the state of the rectified flow at time $t$, and $v(\cdot, t)$ is a trainable velocity field. We seek $v$ that best aligns with the direction of linear interpolation paths. Therefore, the training objective minimizes the discrepancy between the true instantaneous velocity of the interpolation and the learned velocity:

$$\min_{v} \int_0^1 \mathbb{E}\Big[\|(X_1 - X_0) - v_\theta(X_t, t)\|^2\Big]\, dt. \tag{5.3}$$

In practice, we approximate the integral via Monte Carlo sampling:

$$\min_{\theta} \;\; \mathbb{E}_{t\sim \mathrm{U}[0,1]}\Big[\big\|(X_1 - X_0) - v_\theta\big(X_t,\ t\big)\big\|^2\Big]. \tag{5.4}$$

Moreover, the *reflow* procedure, an iterative application of rectified flow training to the outputs of a previous flow, can further straighten trajectories and minimize discretization errors. This

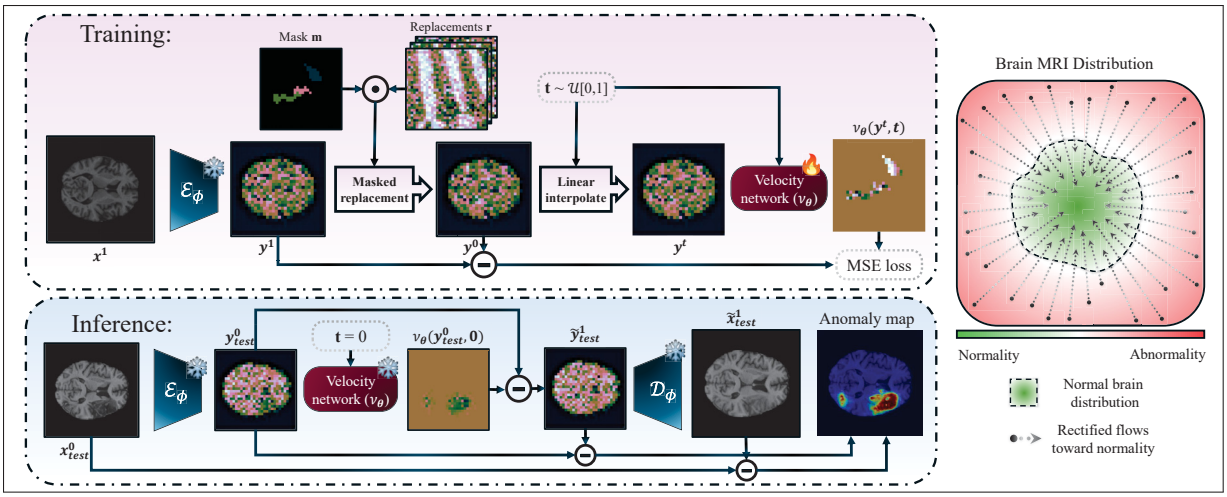



Figure 5.1 **Overview of REFLECT**. *Training:* a velocity network is trained to predict displacement between the latent representation of normal images and their corrupted versions, using their linear interpolation and time $t$. *Inference:* Given an anomalous test image, the flow is obtained using the velocity network, and the input image is corrected in a single step. *Right:* Rectified flows toward normality in brain's latent space

process helps prevent trajectory crossings induced by paired input samples and further refines and rectifies the transport paths toward the target distribution, thereby enhancing robustness and enabling high-quality single-step sampling while preserving distribution fidelity. This is achieved through the following objective:

$$\min_{\theta} \; \mathbb{E}_{Z_0 \sim \pi_0, t \sim \mathrm{U}(0,1)} \left[ \left\| (Z_1 - Z_0) - v'_\theta(Z_t, t) \right\|^2 \right], \quad Z_0 = X_0, \tag{5.5}$$

where $Z_1$ represents the transported sample ($Z_0$) with the initial rectified flow.

## 5.3 Method

Interpreting $\pi_0$ as the entire brain distribution encompassing both normal and abnormal (or injured) brains, and $\pi_1$ as the distribution of normal brains (with $\pi_1 \subset \pi_0$), we train a rectified flow to learn a velocity field $v_\theta(., t)$ that transports abnormal samples from $\pi_0$ toward $\pi_1$ (illustrated in Fig. 5.1), while leaving normal samples unchanged since they already belong to $\pi_1$. Our framework for unsupervised brain anomaly detection using rectified flows (see Figure

5.1) consists of three main stages: (1) generating paired abnormal and normal images in the latent space, (2) rectified flow training to transport abnormal brains toward normality, and (3) localizing anomalies. Below, we describe each stage in detail.

### 5.3.1 Generating Paired Samples

Following (Esser *et al.*, 2024; Beizaee *et al.*, 2025), we transform the data into a latent space to stabilize training, improve output quality, and enable the interpretation of anomalies as noise or out-of-distribution features. To achieve this, we adapt and fine-tune a pre-trained VAE (Rombach *et al.*, 2022) for medical images, which compresses high-dimensional data into a compact latent representation while retaining structures and semantic information. Let $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ denote the set of normal brain samples; the encoder $\mathcal{E}_\phi$ maps each image to a latent code via $\mathbf{y}^{(i)} = \mathcal{E}_\phi(\mathbf{x}^{(i)})$.

To train the rectified flow for transforming anomalous latent representations to normal ones, we construct paired latent vectors $(\mathbf{y}_i^0, \mathbf{y}_i^1)$, where $\mathbf{y}_i^1$ is a normal sample and $\mathbf{y}_i^0$ is an artificially corrupted version. The corruption is applied using a binary mask $\mathbf{m}$ that specifies the regions to be altered, a random replacement vector $\mathbf{r}$ indicating the direction of corruption, and parameter $\alpha \in [0, 1]$ which controls the severity of corruption. The corrupted latent vector is given by:

$$\boldsymbol{y}_i^0 \;=\; \left(\sqrt{1-\alpha}\cdot \boldsymbol{y}_i^1 \;+\; \sqrt{\alpha}\cdot \mathbf{r}\right) \odot \mathbf{m} \;+\; \boldsymbol{y}_i^1 \odot (1-\mathbf{m}), \tag{5.6}$$

which guarantees that if both $\mathbf{y}_i^1$ and $\mathbf{r}$ follow a standard normal distribution, the resulting corrupted image will also follow a standard normal distribution.

The introduced corruptions must closely mimic genuine injuries and span the entire range of possible anomalies. We assume that the brain may exhibit up to $N$ distinct anomalous regions delineated by non-overlapping masks. The mask of each region is generated by a random walk starting from a random point within the brain and taking a random number of uniformly probable steps to neighboring points. Masked positions then correspond to the points visited during the walk. This strategy results in more realistic anomaly shapes than rectangular patch-based masking (Beizaee *et al.*, 2025).

Afterward, a replacement vector is assigned to each masked region, and the final corrupted image is generated according to Eq. 5.6. For each region, the replacement vector is generated by randomly choosing between two distinct strategies. The first strategy is to use random noise, motivated by the interpretation of anomalies as noise in the latent space. Alternatively, we can use a cropped segment from the latent representation of a textured image (not necessarily medical images) to impose more realistic and structured anomalies. For the first strategy, in our implementation, we propose sampling random noise independently for each channel according to:

$$\mathbf{r} = \sqrt{\beta} \cdot q + \sqrt{1-\beta} \cdot \mathbf{p}, \tag{5.7}$$

where $q \in \mathbb{R} \sim \mathcal{N}(0,1)$, $\mathbf{p} \in \mathbb{R}^{H \times W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\beta$ is uniformly sampled from $[0,1]$ (fixed for all channels). This formulation can be interpreted as a weighted combination of a global (image-level) random variable and a spatially varying (pixel-wise) random vector, constructed such that the resulting $\mathbf{r}$ follows a standard normal distribution. This strategy imposes spatial dependencies and yields more realistic and structured corruptions in the latent space.

### 5.3.2 Training Rectified Flows for Anomaly Correction

Our goal is to learn a mapping that transports abnormal latent samples $\boldsymbol{y}^0 \sim \pi_0$ back to the normal distribution $\pi_1$. For this purpose, we train the velocity field $v_\theta(., t)$ using the paired latent data $(\boldsymbol{y}^0, \boldsymbol{y}^1)$. The training objective becomes:

$$\min_{\theta} \ \mathbb{E}_{t \sim \mathrm{U}[0,1]} \Big[ \big\| (Y_1 - Y_0) - v_\theta(Y_t,\ t) \big\|^2 \Big], \tag{5.8}$$

where $Y_t = (1-t)\, Y_0 + t\, Y_1$. Once the velocity model is trained, the learned ODE (Eq. 5.2) defines a flow that can be numerically simulated using a standard ODE solver (e.g., Euler method). Note that in many cases, including our application, a coarse discretization (even a single Euler step) is sufficient due to the straightening effect of the learned flow. Furthermore, after obtaining the first velocity model, we can refine the flow by training a second rectified flow model $v'(.,\, t)$ using the reflow process. This additional stage could further straighten the

trajectories and enhance the correction procedure. We refer to this model as 2-REFLECT, while the first model is denoted as 1-REFLECT.

### 5.3.3 Inference and Anomaly Localization

At test time, given an image $\boldsymbol{x}^0_{\text{test}}$ (potentially containing anomalies), we first encode it as $\boldsymbol{y}^0_{\text{test}} = \mathcal{E}_\phi(\boldsymbol{x}^0_{\text{test}})$. Then, we solve the reverse-time ODE with the trained velocity model using a single Euler step which results in:

$$\tilde{\boldsymbol{y}}^1_{\text{test}} \;=\; \boldsymbol{y}^0_{\text{test}} - \; v_\theta(\boldsymbol{y}^0_{\text{test}},\, 0), \tag{5.9}$$

where $\tilde{\boldsymbol{y}}^1_{\text{test}}$ should ideally be the corrected sample within $\pi_1$ distribution. The reconstructed normal image is decoded from the corrected latent space using the VAE decoder: $\tilde{\boldsymbol{x}}^1_{\text{test}} = \mathcal{D}_\phi(\tilde{\boldsymbol{y}}^1_{\text{test}})$. Finally, anomalies are localized by comparing the original image to its reconstruction in both latent space and image space:

$$\mathbf{A}(\mathbf{x}_{\text{test}}) \;=\; \frac{1}{2}\left|\tilde{\boldsymbol{x}}^1_{\text{test}} - \boldsymbol{x}^0_{\text{test}}\right| \;+\; \frac{1}{2}\left|\tilde{\boldsymbol{y}}^1_{\text{test}} - \boldsymbol{y}^0_{\text{test}}\right|. \tag{5.10}$$

Higher differences indicate regions where the flow had to make larger corrections, thereby signaling anomalies.

## 5.4 Experiments

### 5.4.1 Experimental Setting

**Datasets.** For our experiments, we employed BraTS'21 (Bakas *et al.*, 2017) that comprises 1251 brain scans across four modalities (T1, Contrast-Enhanced T1 (T1CE), T2, and FLAIR), and ATLAS 2.0 (Liew *et al.*, 2022), which contains 655 T1-weighted MRI scans. Both datasets come with expert-annotated lesion masks. We extracted 20 central axial slices from skull-stripped brains and padded them to a resolution of 256×256 pixels. Both datasets are divided into training (80%), validation (10%), and testing (10%) subsets, and only normal training slices are used for

Table 5.1 **Quantitative results** obtained by different approaches. The best method per modality and/or dataset is highlighted in **bold**, whereas the second one is underlined. The performance gains over the best baseline are shown in green

| Method | | ATLAS 2.0 | BraTS'21 | | | |
|---|---|---|---|---|---|---|
| | | T1-w | FLAIR | T1CE | T2-w | T1-w |
| AE (Baur *et al.*, 2021a) | *MedIA*'21 | 11.9 | 33.4 | 32.3 | 30.2 | 28.5 |
| DDPM (Ho *et al.*, 2020b) | *Neurips*'20 | 20.2 | 60.7 | 37.9 | 36.4 | 29.4 |
| AutoDDPM (Bercea *et al.*, 2023a) | *ICML Workshop*'23 | 12.7 | 55.5 | 36.9 | 29.7 | 33.5 |
| DAE (Kascenas *et al.*, 2022) | *MIDL*'22 | 11.1 | 79.7 | 36.7 | 69.6 | 29.5 |
| Cycl.UNet (Liang *et al.*, 2023) | *MICCAI*'23 | N/A | 65.0 | 42.6 | 49.5 | 37.0 |
| IterMask$^2$ (Liang *et al.*, 2024a) | *MICCAI*'24 | 35.3 | <u>80.2</u> | 61.7 | 71.2 | 58.5 |
| MAD-AD (Beizaee *et al.*, 2025) | *IPMI*'25 | <u>36.1</u> | 76.2 | <u>68.5</u> | <u>73.2</u> | <u>63.4</u> |
| **1-REFLECT** | *Ours* | **41.6**$_{+5.5}$ | **85.1**$_{+4.9}$ | **73.0**$_{+4.5}$ | 79.6$_{+6.4}$ | 69.7$_{+6.3}$ |
| **2-REFLECT** | *Ours* | 40.8 $_{+4.7}$ | 83.2$_{+3.0}$ | 72.0$_{+3.5}$ | **80.3**$_{+7.1}$ | **69.8**$_{+6.4}$ |

training, while the single slice of the test subjects displaying the most prominent pathology is reserved for inference. Moreover, the Describable Textures Dataset (DTD) (Cimpoi *et al.*, 2014) is used and converted to gray-scale to serve as textured replacement images.

**Evaluation metrics.** Following (Beizaee *et al.*, 2025), we assess models' performance using the Maximum Dice score, which reflects the highest Dice coefficient achieved as the threshold varies from 0 to 1.

**Implementation details.** In each iteration, the number of masked regions is randomly chosen between 1 and 4, and the number of random-walk steps is randomly sampled between 0 and 200. 1-REFLECT model underwent training for 200 epochs for BraTS'21 and for 400 epochs for the ATLAS dataset with a batch size of 96, using the AdamW optimizer and a learning rate of $5 \times 10^{-4}$. Afterward, 2-REFLECT was trained on top of 1-REFLECT for another 50 epochs with a learning rate of $1 \times 10^{-5}$. Also, we have used 5 reverse ODE correction steps.

### 5.4.2 Results

**Main quantitative results.** We empirically evaluate the performance of the proposed approach compared to relevant UAD methods proposed for brain MRI, whose results are reported in Table 5.1. These values demonstrate that REFLECT substantially outperforms recent state-of-the-art

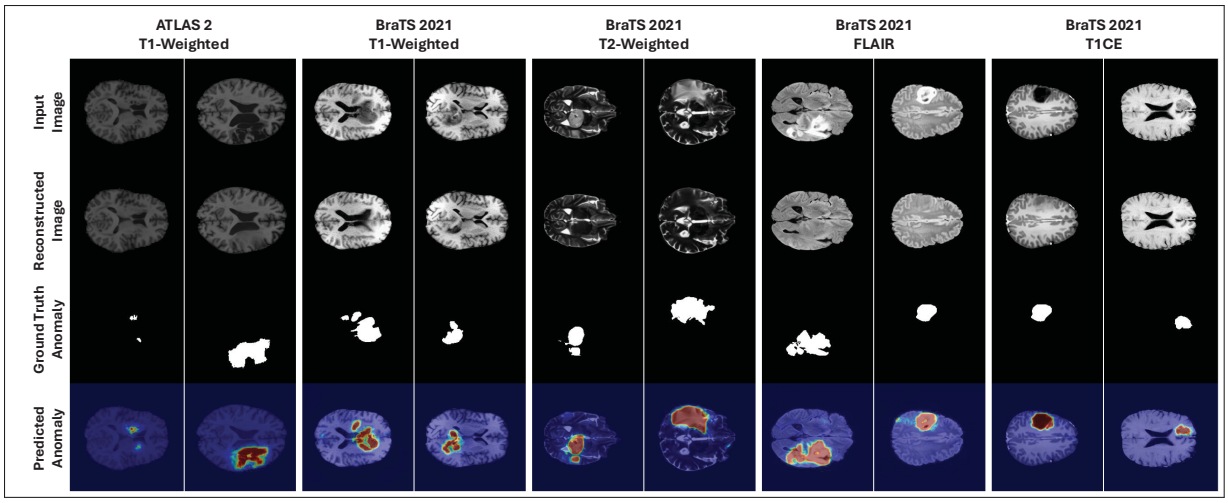



Figure 5.2 **Qualitative results.** *Second row:* Reconstructed images of their abnormal input counterparts. *Last row:* Anomaly segmentation maps obtained by our approach

brain UAD methods. Compared to the second best approach, i.e., MAD-AD (Beizaee *et al.*, 2025), our method brings improvement gains ranging from 4.5% to 6.4% in both datasets. In particular, in T1-w, which seems to be the most difficult modality based on baselines' results, REFLECT yields the highest difference gaps (similar to gains in T2-w). These results demonstrate the superiority of our method across modalities and datasets, compared to very relevant baselines. Regarding **reflow**'s effect (i.e., 2-REFLECT) on the first trained model 1-REFLECT, results (Table 5.1) suggest that the reflow process did not improve the overall performance of the anomaly detection, likely because the initial flows are already well-rectified, and the reflow process is attached to the anomaly localization performance of the first method.

**Qualitative results.** To visually assess the effectiveness of REFLECT, we depict in Fig. 5.2 several visual examples of both the reconstructed images (*second row*) and the predicted anomaly maps (*last row*) across all modalities of the BraTS dataset. As these images highlight, REFLECT successfully reconstruct realistic "healthy" images, which results in accurate predicted anomaly maps. Fig. 5.3 depicts a visual example of the transitions performed by the rectified flow.

**Ablations.** *Model size:* We evaluated five different model sizes to analyze the trade-off between model efficiency and anomaly detection performance. While larger models demonstrated slightly

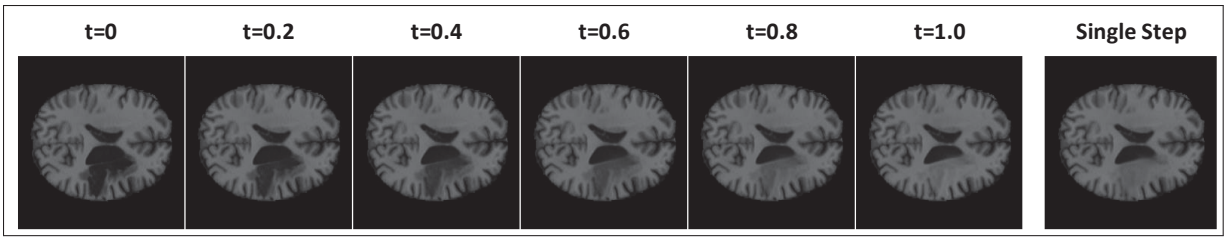



Figure 5.3 Transition between an anomalous brain to its healthy counterpart using 10 reverse ODE Euler steps *vs.* a single step, demonstrating correction flows are well-rectified, and can correct abnormalities in a single step

Table 5.2 Ablation studies (BRATS'21) on model size and VAE model employed

| | Model | #Params | Modality | | | | |
|---|---|---|---|---|---|---|---|
| | | | FLAIR | T1CE | T2-w | T1-w | Avg |
| Model size | UNet XS | ~16 M | 81.7 | 70.6 | 80.9 | 69.7 | 75.7 |
| | UNet S | ~64 M | 83.9 | 72.1 | 81.0 | 69.6 | 76.7 |
| | UNet M* | ~145 M | **85.1** | 73.0 | 79.6 | 69.7 | 76.9 |
| | UNet L | ~257 M | 83.2 | 72.1 | **80.8** | **71.2** | 76.8 |
| | UNet XL | ~580 M | 83.6 | **73.3** | 79.9 | 70.7 | **76.9** |
| VAE | KL-f4 | ~55 M | **85.4** | 72.1 | **80.8** | **71.8** | **77.5** |
| | KL-f8* | ~84 M | 85.1 | **73.0** | 79.6 | 69.7 | 76.9 |

better performance, smaller variants still achieved highly competitive results, significantly outperforming previous methods, making REFLECT a strong choice for real-time or resource-constrained applications. *Effect of the VAE model:* We also investigated the effect of the VAE model by comparing two variants with scale factors of 4 and 8. As illustrated in Table 5.2, both VAEs perform well, with the VAE using a scale factor of 4 exhibiting slightly better performance, likely due to its higher embedding dimensionality. However, this comes at the cost of increased computational requirements.

## 5.5 Conclusion

We introduced REFLECT, an unsupervised framework that leverages rectified flows in latent space, which provides a direct one-step correction transport that significantly improves reconstruction

fidelity and anomaly localization. Experimental results on established unsupervised anomaly detection benchmarks confirm that REFLECT outperforms current state-of-the-art approaches, paving the way for more robust and efficient diagnostic tools in neuroimaging. Future directions include extending the framework to 3D data and testing cross-dataset generalization to boost clinical relevance.

# CHAPTER 6

# DETERMINING REGIONAL BRAIN GROWTH IN PREMATURE AND MATURE INFANTS IN RELATION TO AGE AT MRI USING DEEP NEURAL NETWORKS

Farzad Beizaee[1,2] , Michele Bona[1,2] , Christian Desrosiers[1] , José Dolz[1] , Gregory Lodygensky[2,3]

[1] ÉTS Montreal, Canada
[2] CHU Sainte-Justine, University of Montreal, Canada
[3] Canadian Neonatal Brain Platform, Montreal, Canada.

Article published in the journal "Scientific Reports", August 2023

## 6.1 Introduction

MRI is increasingly used in neonates as it provides a wealth of information vastly superior to ultrasound and CT scans. Simple preparation with sedation achieved by milk alone and swaddling is sufficient to guarantee high-quality, motionless imaging data without needing any anesthetics (Mathur *et al.*, 2008). However, it can be challenging to analyze neonatal brain MRIs due to the lack of readily-available and age-specific references. Having a fast and simple tool that assesses brain maturation during such a period of extraordinary changes would be immensely helpful. Brain segmentation using standard image analysis tools has been tremendously useful in analyzing brain development in the last few decades. Already in 1998, Hüppi *et al.* (1998) used the k-nearest-neighbor (k-NN) classification to show how gray matter volumes correlated significantly with postmenstrual age at MRI, more so than unmyelinated white matter.

Since then, more advanced analyses studying cortical folding in preterm infants have shown how this process is tightly controlled during the last trimester with a strong correlation to postmenstrual age. As described by Dubois *et al.* (2008), the general proportion of sulci compared to the brain size was found to correlate to the postmenstrual age of the infant. Shimony *et al.* (2016) also demonstrated that the general curvature and sulcal depth of the brain were highly correlated with age. More recently, Galdi *et al.* (2020) segmented the brain into 81 regions and then extracted related features from structural MRI and diffusion MRI, which served

to predict postmenstrual age (PMA) at scan based on inter-regional similarities. The advantage of these techniques is that each step can be visualized and validated. On the other hand, their use in daily clinical practice is not viable due to the many processing steps and high computing time.

In recent years, deep learning-based models have demonstrated an astonishing performance on a wide range of medical problems such as classification, detection, and segmentation (Litjens *et al.*, 2017; Esteva *et al.*, 2021). Indeed, tools based on deep learning are readily available for both research and daily clinical practice. Once properly trained and validated, they can be installed on practically any computer and provide holistic quantitative information in a matter of minutes. Furthermore, having a learning methodology based solely on structural T2-weighted images would guarantee its generalizability, as any center with access to an MRI machine would be able to secure data with sufficient quality for interpretation. Widely-used in other medical research fields, deep neural networks (DNN) have also found their way to the task of brain estimation using medical images. Recently, several brain age estimation approaches based on deep learning were proposed for brain age estimation in fetuses, infants, and adults. Peng *et al.* (2021) proposed to use a simple fully convolutional network for this task, arguing that it does not require a very deep neural network. Furthermore, they advocated that smaller networks would be more robust for small datasets. In their study, Cheng *et al.* (2021) leveraged a ranking loss term and a two-stage cascaded 3D convolutional neural network (CNN) to improve the accuracy of brain age estimation. A transformer-shaped network was proposed by He, Grant & Ou (2021) to fuse local features from smaller patches with global features by using an attention mechanism. Shi *et al.* (2020) employed an attention-based deep residual network with structural MRI to predict the brain age. They also computed the predictive uncertainty using an ensembling strategy to estimate the model's confidence as a marker for fetal brain anomaly detection. Hong *et al.* (2021) employed multiplanar slices in orthogonal directions and a test-time-augmentation technique to predict the brain age based on each slide. The most frequent value among the predictions was used as the estimated age. More recently, Taoudi-Benchekroun *et al.* (2022) exploited Diffusion and T2-weighted MRI images to generate individual brain connectivity maps, which were later used by a deep network to predict age.

Despite the growing interest in brain age estimation using DNNs, yet due to the immense power of these networks to handle a vast quantity of data, they tend to resort to multimodal data, which is not always accessible, or too complex deep neural networks that might limit their reproducibility and highly increase their computation resources. Another major deterrent of DNN for radiologists is their lack of interpretability. Indeed, it is vital that the structures driving the results are detailed and compared with available literature. Furthermore, identifying important features is of paramount importance in case of brain injury, where regional changes could greatly affect the accuracy of results. Unfortunately, existing approaches based on deep models tend not to provide a mechanism to identify the most important structures for the age estimation task.

Motivated by the aforementioned limitations, we designed a learning-based pipeline to predict neonatal brain age using T2-weighted MRIs. The proposed approach first segments the T2-weighted MRIs into 87 cortical and sub-cortical structures, which are then used to extract features associated with volume and gyrification. These features, termed relational volume and surface to volume ratio, are informative for brain age estimation yet are easy to calculate. Finally, a machine learning regression model can be trained to predict brain age using the extracted features from the segmented MRIs. The overview of the proposed pipeline is shown in Figure 6.1. The proposed pipeline predicts brain age in a fast and accurate manner. It is also interpretable, as we can identify the most important features driving the results using popular features selection techniques, such as Permutation Feature Importance (PFI) (Breiman, 2001). It is important to note that the term "interpretability" in our context refers to the ability to extract and understand the significant factors and features utilized by the model in predicting brain age using MRI data. This extracted information holds the potential for examination and validation through surveys conducted with end-users, such as radiologists. Additionally, it can serve as a valuable resource in guiding future research focused on brain development during the last trimester or identifying landmarks for neonatal brain age.

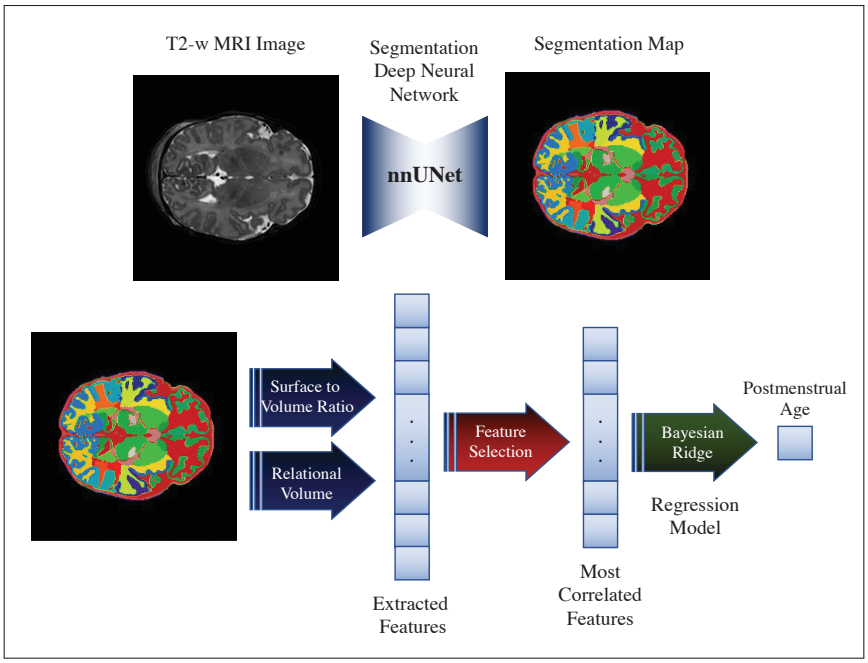



Figure 6.1 Overview of the proposed method. First, we employ 3D T2 weighted MRI as input for the segmentation model. The generated segmentations were then used to extract the volume and surface of each class (i.e., 87 classes representing 87 brain regions). Using these measures, we calculated the proposed metrics, i.e., the surface to volume ratio and the relational volume. After this step, we resorted to a feature selection strategy, known as PFI, to keep the most important regions for brain age estimation. Finally, we used a Bayesian ridge regression model to get the predicted postmenstrual age

## 6.2 Results

### 6.2.1 Main results

We used the developing Human Connectome Project (dHCP)(Makropoulos *et al.*, 2018) dataset to empirically validate the proposed method. Following standard practices in machine learning, the data was split into three distinct random subsets, resulting in independent train, validation and test sets. The train data, which includes 60% of the scans (334 images), was first used to train the segmentation and regression models for brain age estimation. Then, the validation data,

which includes 15% of the data (84 images), was employed to find the optimal parameters and select the best models. Last, the remaining 25% of the data (140 images) was used for testing the method and comparing to other approaches.

The quantitative results achieved by the proposed framework for brain age estimation are reported in Table 6.1. Recent works that solely employ structural MRI for the same task are also included for comparison purposes. From these results, we observe that the proposed approach largely outperforms recent literature in terms of the Mean Absolute Error (MAE), achieving a value of 0.46 weeks on the independent test set. Compared to relevant works (2DTTA, ARN, and GLT), our approach brings between 5 and 11% improvement. A noteworthy point to highlight is the fact that existing approaches predict brain age directly from the structural MRI, which makes the interpretation of the predictions difficult. In contrast, as our method employs the segmentation results to derive two per-region features, these can be selected according to their correlation with brain and gestational age. As we show later in our empirical validation, this allows shedding light on the regions that have a significant impact on brain age estimation.

Table 6.1 Quantitative results compared to state-of-the-art learning-based brain age estimation methods

| model | $MAE$ | $R^2$ | $RMSE$ |
|---|---|---|---|
| 2D Test-time-Augmentation (2DTTA) (Hong *et al.*, 2021) | 0.52±0.40 | 0.96 | 0.66 |
| Attention-based residual network (ARN) (Shi *et al.*, 2020) | 0.57±0.44 | 0.95 | 0.75 |
| Global-local transformer (GLT) (He *et al.*, 2021) | 0.51±0.39 | 0.96 | 0.64 |
| **Proposed method** | **0.46** ±0.37 | **0.97** | **0.60** |

Figure 6.2 depicts additional results obtained by our method. Figure 6.2a shows the relation between the predicted *versus* the real age, which shows a high correlation between both. Indeed, these visual results are supported by the high determination score ($R^2$) of 0.97 obtained by the proposed approach. Despite differences in the mean absolute error across age intervals (Figure 6.2b), we can see that these are typically consistent, with slightly lower means as the age increases. Last, we observe that the MAE obtained by our approach for the “male” and “female” populations are quite similar (0.46 and 0.47 weeks, respectively), indicating that our method is gender agnostic.

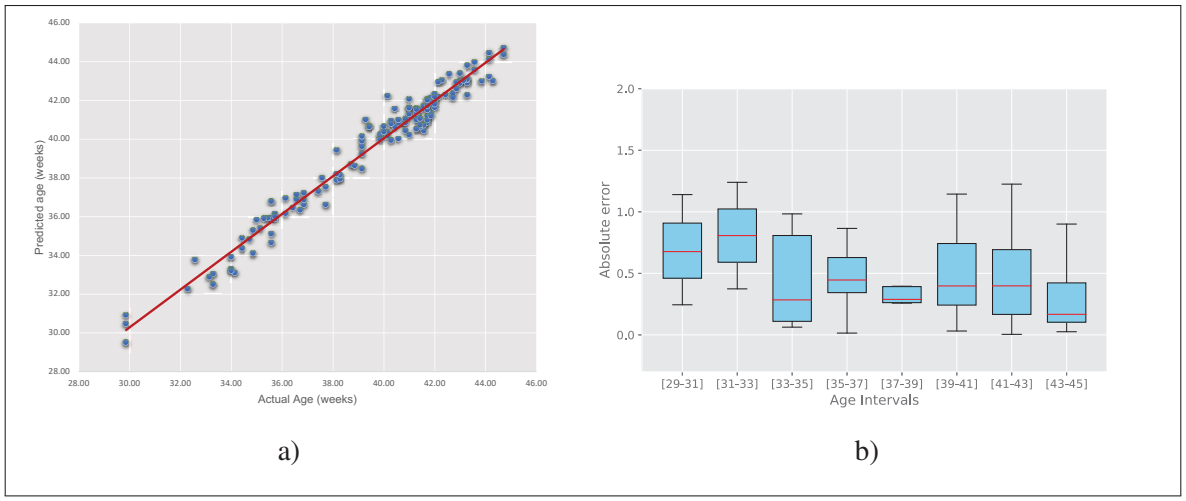



Figure 6.2 a) Predicted age *versus* real age; b) Prediction error for different age intervals

### 6.2.2 On the importance of different regions

One of the advantages of the proposed method is that it is interpretable in terms of the importance of each region and its corresponding features. Therefore, it is possible to extract information from the trained regression model, which provides a possible landmark to look for predicting brain age. To understand the importance of each region-feature pair in brain age estimation, two regression models are trained with the surface to volume ratio (SVR) and relational volume (RV) separately. Then, the importance of each region for each of the features is calculated using a permutation feature importance approach. Finally, the importance of each region is re-scaled to be constrained between 0 and 1, for better interpretability. These observations can then be used as a valid biomarker for neonatal brain age estimation. The most important regions and their importance for relational volume and surface to volume ratio are shown in Figure 6.3. Based on the obtained regions' importance, the extracted relational volumes (RV) of frontal lobe gray and white matter and parietal gray matter have the most correlation with neonatal brain age. The scaled importance factors for these structures are higher than 0.4, and the remaining structures have lower scaled importance. Additionally, we can also observe that frontal and parietal lobe white matter and thalamus surface to volume ratios (SVR) are the most informative regions for predicting brain age, with scaled importance factors of more than 0.6.

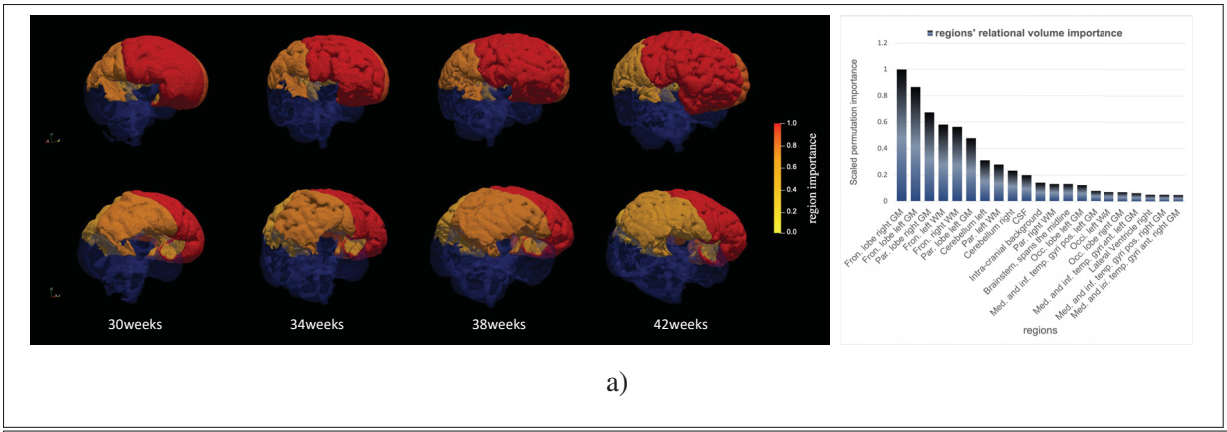



Figure 6.3 Histogram showing the brain regions with the best scaled PFI (on the right) and a 3D representation of these regions colored by their PFI value (on the left). The brains are selected from the dHCP dataset aged (from left to right) 30, 34, 38, and 42 weeks. The first row (a) shows that the most significant brain structures for RV were the frontal lobe right and left gray matter (Right PFI 0.1, Left PFI: 0.86), the parietal lobe right gray matter (PFI: 0.67), the frontal lobe left and right white matter (Left PFI: 0.58, Right PFI 0.56) then the parietal lobe left gray matter (PFI: 0.48). The second row (b) shows that the most significant brain structures for SVR were the frontal lobe right and left white matter (Right PFI 1.0, Left PFI: 0.97), the parietal lobe right white matter (PFI: 0.9), then the right and left thalamus (Right PFI 0.8, Left PFI: 0.71)

### 6.2.3 Comparison to existing brain age bio-markers

In this section, we compared the two proposed features, i.e., Relational Volume (RV) and Surface to Volume Ratio (SVR) to existing biomarkers commonly used to estimate brain age: sulcal depth, cortex thickness, and curvature. These experiments, whose results are reported in Table 6.2, demonstrate that the proposed features indeed outperform other proposed metrics in the

literature. In particular, compared to the cortex thickness and curvature, the improvements brought by the proposed features are substantial.

Table 6.2 Comparison of the quantitative performance obtained by the proposed features and existing features for brain age estimation

| feature | $MAE$ | $R^2$ | $RMSE$ |
|---|---|---|---|
| Sulcal depth | 0.67 ±0.52 | 0.93 | 0.85 |
| Cortex thickness | 1.37 ±1.08 | 0.72 | 1.75 |
| Curvature | 1.02 ±0.76 | 0.85 | 1.27 |
| RV and SVR **(Proposed)** | **0.50 ±0.38** | **0.96** | **0.62** |

### 6.2.4 Performance on low labeled data regime

It is well known that deep learning models require large amounts of training labeled data to work satisfactorily. Nevertheless, due to several factors, including time-consuming processes and annotator variability, having access to such large curated datasets is not always easy. Thus, we believe that it is important to investigate the effect of low labeled data regime on the performance of different approaches. To this end, we repeated the training of several approaches under several amounts of labeled training data, and reported their results in Table 6.3. An interesting observation is that, particularly in the most extreme scenario (i.e., only 10 labeled scans were used for training), the performance gap between the proposed method and existing literature is substantially large. More concretely, recent works achieve an MAE ranging from 1.37 to 2.39 weeks, whereas our method can predict the brain age with an MAE of 0.74 weeks, which represents less than half of the value obtained by compared approaches. We also observed that even though the difference between the different approaches is reduced as the number of labeled samples increases.

We repeated the training of the segmentation and regression models with only a fraction of data to analyze the amount of data required for training and its effects on segmentation and brain age estimation. It is important to investigate the effect of a low data regime since medical data is limited, costly, and hard to obtain. Also, the performance of the model based on the

low data shows its generalizability and robustness, which is a crucial factor for medical image analysis. For the low data regime, each time, 10, 20, 40, 60, and 100 scans were selected from the training split to train the models respectively. Also, 40% of the training data was used as the validation set. However, the models were tested with the whole test split (140 scans). Besides, we preserved the same images for training and validation sets of different experiments for a fair comparison. We also tried this configuration for all comparing methods for better analysis.

Table 6.3 Quantitative performance of different brain age estimation methods based on a fraction of labeled data

| # Training images | Method | $DSC$ | $MAE$ | $R^2$ | $RMSE$ |
|---|---|---|---|---|---|
| 10 scans | 2DTTA (Hong *et al.*, 2021) | - | 1.37±1.21 | 0.70 | 1.82 |
| | ARN (Shi *et al.*, 2020) | - | 2.39±1.52 | 0.14 | 2.7 |
| | GLT (He *et al.*, 2021) | - | 1.77±1.30 | 0.57 | 2.2 |
| | **Proposed** | 0.846 | **0.74 ±0.57** | **0.92** | **0.94** |
| 20 scans | 2DTTA (Hong *et al.*, 2021) | - | 0.84±0.78 | 0.88 | 1.15 |
| | ARN (Shi *et al.*, 2020) | - | 1.39±1.13 | 0.70 | 1.57 |
| | GLT (He *et al.*, 2021) | - | 1.09±0.80 | 0.83 | 1.36 |
| | **Proposed** | 0.875 | **0.68 ±0.54** | **0.93** | **0.86** |
| 40 scans | 2DTTA (Hong *et al.*, 2021) | - | 0.70±0.64 | 0.92 | 0.95 |
| | ARN (Shi *et al.*, 2020) | - | 0.96±0.83 | 0.86 | 1.14 |
| | GLT (He *et al.*, 2021) | - | 0.80±0.68 | 0.90 | 1.05 |
| | **Proposed** | 0.895 | **0.57 ±0.45** | **0.95** | **0.73** |
| 60 scans | 2DTTA (Hong *et al.*, 2021) | - | 0.61±0.50 | 0.95 | 0.79 |
| | ARN (Shi *et al.*, 2020) | - | 0.96±0.79 | 0.87 | 1.11 |
| | GLT (He *et al.*, 2021) | - | 0.73±0.54 | 0.93 | 0.91 |
| | **Proposed** | 0.901 | **0.56 ±0.40** | **0.95** | **0.68** |
| 100 scans | 2DTTA (Hong *et al.*, 2021) | - | 0.60±0.47 | 0.95 | 0.76 |
| | ARN (Shi *et al.*, 2020) | - | 0.76±0.59 | 0.92 | 0.93 |
| | GLT (He *et al.*, 2021) | - | 0.58±0.44 | 0.95 | 0.73 |
| | **Proposed** | 0.909 | **0.51 ±0.39** | **0.96** | **0.64** |

### 6.2.5 The impact of different backbones

The choice of different segmentation or regression models can arguably have a significant impact on the final performance. To assess these potential performance differences, we investigate several segmentation and regression models in our framework, whose results are reported in Table 6.4.

In terms of the segmentation model, we first evaluated the performance of the brain labels extracted by DRAW-EM (Developing brain Region Annotation With Expectation-Maximization) (Makropoulos *et al.*, 2014, 2016), a popular software for neonatal brain MR image segmentation. Furthermore, we included two well-known medical image segmentation networks (UNet (Ronneberger *et al.*, 2015) and nnUNet (Isensee *et al.*, 2021)), which achieve state-of-the-art results in a broad span of medical segmentation problems. Note that both UNet and nnUNet were trained with the labels generated by DRAW-EM. From the results in Table 6.4, *top* we can observe that despite the segmentation results might be different across networks nnUNet achieved 0.920 of DSC *versus* 0.907 of UNet, the MAE values obtained without feature selection are almost identical. Nevertheless, when the most correlated features were selected, the performance differences between nnUNet and UNet are larger. This indicates that the feature selection process indeed plays an important role in removing noise from uncorrelated, or not as much correlated features. Furthermore, regardless of the segmentation method employed, the obtained results outperform the current literature, whose achieved MAE results were 0.51 (GLT), 0.52 (2DTTA), and 0.57 (ARN). An interesting and surprising observation is that, while UNet and nnUNet were trained from DRAW-EM segmentation labels, they typically yield better results. Thus, these results indicate that even though the brain age estimation accuracy increases with the segmentation accuracy, the proposed method does not require very complex segmentation networks to achieve satisfactory performances, which contrasts with prior works.

We now evaluate the effect of different regression models in Table 6.4, *bottom*. In contrast to the previous observations regarding the segmentation model, the choice of the regression approach significantly impacts brain age prediction accuracy. In particular, the differences between ElasticNet (worst) and Bayesian Ridge (best) as regression approaches are equal to 0.32 and 0.36 weeks without and with feature selection, respectively. This indicates that even though our pipeline is sensitive to the choice of the regression method, the achieved results by most models can be considered satisfactory and shows that our pipeline is model-agnostic. However, we suggest that a proper validation must be conducted to select the best strategy. Similar to the segmentation scenario, the step of feature selection further improves brain age prediction, which

demonstrates its usefulness in finding correlated features and removing potential sources of noise during the learning process.

Table 6.4 Results using different segmentation and regression backbones

| Method | Without feature-selection | | | With feature-selection | | |
|---|---|---|---|---|---|---|
| | $MAE$ | $R^2$ | $RMSE$ | $MAE$ | $R^2$ | $RMSE$ |
| Segmentation models | | | | | | |
| DRAW-EM (Makropoulos *et al.*, 2014) | 0.51 ±0.40 | 0.96 | 0.65 | 0.50 ±0.39 | 0.96 | 0.63 |
| UNet (Ronneberger *et al.*, 2015) | 0.52 ±0.40 | 0.96 | 0.65 | 0.50 ±0.40 | 0.96 | 0.63 |
| **nnUNet** (Isensee *et al.*, 2021) | **0.50 ±0.38** | **0.96** | **0.62** | **0.46 ±0.37** | **0.97** | **0.60** |
| Regression models | | | | | | |
| Kernel Ridge (Murphy, 2012) | 0.61 ±0.50 | 0.94 | 0.79 | 0.56 ±0.44 | 0.95 | 0.72 |
| ElasticNet (Zou & Hastie, 2005) | 0.82 ±0.55 | 0.91 | 0.98 | 0.82 ±0.55 | 0.91 | 0.98 |
| GradientBoosting Friedman (2002) | 0.57 ±0.48 | 0.95 | 0.73 | 0.58 ±0.49 | 0.95 | 0.76 |
| SVM Regressor (Drucker *et al.*, 1996) | 0.75 ±1.00 | 0.86 | 1.25 | 0.61 ±0.81 | 0.91 | 1.01 |
| MLP Regressor (Hinton, 1990) | 0.58 ±0.41 | 0.95 | 0.76 | 0.55 ±0.45 | 0.95 | 0.72 |
| **Bayesian Ridge** (Tipping, 2001) | **0.50 ±0.38** | **0.96** | **0.62** | **0.46 ±0.37** | **0.97** | **0.60** |

## 6.3 Discussion

Relational volume (RV) and surface to volume ratio (SVR) in 87 cortical and sub-cortical brain classes were extracted using a fully automated pipeline built on a combination of machine learning approaches to determine postmenstrual age at MRI, which achieved an MAE of 0.46 weeks. For this purpose, the dHCP database(Makropoulos *et al.*, 2018) has been used which includes 558 neonatal brain T2-weighted MRIs ranging from 29 to 45 weeks coupled with brain regions' contours extracted from DrawEM(Makropoulos *et al.*, 2014). First, using 60% and 15% of the scans as training set and validation set respectively, a segmentation network (nnUNet(Isensee *et al.*, 2021)) was trained to segment the brain MRIs into 87 different regions. Then using the segmented regions, RV and SVR features, which are easy to calculate, are obtained for each region and the ones with the highest correlation with brain age are selected to train a regression model. Once the model was trained, we evaluated the proposed pipeline for neonatal brain age estimation using the remaining 25% of scans. Finally, we wanted to assess critical structures and morphometric features indicative of brain age estimation, and using PFI

we found that frontal and parietal lobes and thalami were among the most important structures driving these results in a cohort of healthy term and preterm infants.

Adding the proposed measurements brought a substantial improvement compared to more established biomarkers, such as cortex thickness or curvature. We also evaluated several regression networks and showed the advantage of the Bayesian Ridge approach. We believe that the benefit of using it may be linked to its potential to handle limited data (Tipping, 2001). Still, it is essential to note that the results remain very good for most of the regression models showing that the pipeline is not entirely dependent on a specific regression model (Table 6.4). Furthermore, we have assessed the effect of having different segmentation networks in the first stage of the proposed pipeline, and despite differences in the segmentation results, the choice of this backbone does not have a significant impact on the final brain age estimation task.

The use of permutation feature importance unveiled critical structures and morphometric features used to determine brain age between 29 and 45 weeks postmenstrual age at scan (Figure 6.3). The most significant structures when using the relation volume (RV) feature were both frontal gray and white matter lobes, parietal gray matter lobes, parietal left white matter, and the cerebellum. These findings are well aligned with the study by Gui *et al.* (2019), showing how cortical gray matter and cerebellum using conventional metrics had the fastest growth structures during the period of prematurity. Furthermore, Hong *et al.* (2021) employed a saliency visualization method in fetal MRIs and also found that the cortex and ventricles were major regions for age estimation. Unfortunately, cortical parcellation was not available in both studies, so comparison with the frontal and parietal lobes is not possible.

The most significant morphometric features by quantifying the surface to volume ratio (SVR) were frontal and parietal white matter and both thalami (Figure 6.3). The importance of thalamic maturation has already been described by Deprez *et al.* (2018) with a strong age estimation potential shown by root mean squared errors (RMSEs) of 1.41 weeks in newborns between 29 and 44 weeks. Interestingly, it is not frontal and parietal cortical but white matter folding that were the most significant morphometric features. This might be explained by cortical thickness

in relation to image acquisition parameters with significant partial volume effects between gray matter and cerebro-spinal fluid. Indeed the boundary between gray matter and white matter was preferred in several prior works in this age range (Hill *et al.*, 2010a; Engelhardt *et al.*, 2015). Similarly, using conventional image analysis tools major gyrification growth has been identified in the frontal lobe ($R^2$=0.84) with an even higher correlation in the Temporal-parietal-occipital region ($R^2$=0.9). Hill *et al.* (2010b) described a similar major cortical expansion in frontal and parietal and temporal relative to others when comparing newborns to young adults.

Last, we would like to highlight that, for the sake of fairness, we have conducted experiments not only on the proposed approach but also on three recent relevant works. In particular, the four evaluated methods are trained and evaluated under the exact same conditions and with the same patients, which makes the results across models directly comparable. Under this scenario, the proposed pipeline outperformed recent methods, setting new state-of-the-art results for the task of image-based brain age estimation using only a single MRI modality. Last, it is noteworthy to mention that there exist other methods that were not included in the empirical validation, mainly due to different settings, e.g., multimodal images (Galdi *et al.*, 2020; Taoudi-Benchekroun *et al.*, 2022). Nevertheless, the results obtained by these approaches were far from the performance achieved by our pipeline. For example, Liu *et al.* (2024) evaluated several structures and obtained a maximum MAE of 1.19 weeks using T1-weighted images alone. Galdi *et al.* (2020) obtained an MAE of 0.70 weeks in newborns scanned between 38 and 45 weeks postmenstrual age based on multimodal data from T1 and T2-weighted imaging and Multi-shell diffusion MRI using a linear regression model with elastic net regularization. Taoudi-Benchekroun *et al.* (2022) achieved similar results with an MAE of 0.72 weeks in newborns scanned from 37 to 45 weeks from the dHCP cohort combining multimodal data extracted from T2-weighted and Multi-shell diffusion MRI; While our method achieved a better MAE (0.46) despite of considering a wider age range.

### 6.3.1 Limitations

This study characterized a healthy cohort of preterm and term infants. The results may be different in preterm infants with selective injury. The interesting aspect of PFI is that we now have identified critical structural and morphometric features in the neonatal brain driving brain age assessment. It will be critical in the future to assess the reliability of this approach in newborns with brain injury. When studying the impact of prematurity itself on cortical folding, the main structures affected were the insula, superior temporal sulcus, and ventral portions of the pre- and postcentral sulci, features that are very different from the ones we identified to determine brain age. It will be critical before translating these powerful tools into daily clinical practice to determine their efficacy in several clinical situations such as diffusion and cystic white matter injury and that interpretability be always accessible.

Compared to deep learning models trained end-to-end, our method considers feature learning and brain age estimation as two separate steps. A drawback of this strategy is that the learned features may not be optimal for the prediction task. However, it also brings the significant advantage of making the results of method more interpretable, since it enables identifying the regions and imaging features that are more important for brain age estimation.

In this study, we validated our method on the developing Human Connectome Project (dHCP) dataset. However, as it relies on structural characteristics (e.g., volume and gyrification) which are not dataset-dependant, our method could be easily adapted to any other dataset. Furthermore, with recent advancements in domain adaptation and harmonization, segmentation networks trained on given data can be easily adapted to MRIs from other datasets, making our brain age estimation method generalizable.

## 6.4 Materials and Methods

In this section we present the datasets used and detail the experiments and methodology employed in this work.

### 6.4.1 Dataset

To empirically validate the proposed approach for brain age estimation we resorted to the T2-weighted (T2-w) images from the developing Human Connectome Project (dHCP) (Makropoulos *et al.*, 2018) data. Imaging was carried out on 3T Philips Achieva using a dedicated neonatal imaging system. There are 558 sessions with T2-w images that passed QC from 505 different subjects. Infants from the DHCP database were recruited and imaged at the Evelina Neonatal Imaging Centre, London. Informed parental consent was obtained for imaging and data release, and the study was approved by the UK Health Research Authority (Hughes *et al.*, 2017). The medical ethical review board of CHUSJ hospital approved the study. Also, all the methods were performed in accordance with the relevant guidelines and regulations.

The volumes were segmented into 87 regions using the atlas-based segmentation approach known as DRAW. To provide structural priors, manually labeled atlases with expert neuroanatomist annotations are registered to the volume. Afterward, segmentation is carried out using an Expectation-Maximization technique that combines the volume's intensity model with structural priors (Makropoulos *et al.*, 2014, 2016). Among those 505 subjects, 222 subjects are female and the remaining 283 are male. They are born between $24^{+2}$ weeks and $42^{+2}$ weeks of gestation, and scanned at age from $29^{+2}$ weeks to $45^{+1}$ weeks. The available dHCP dataset was split into three completely distinct sets of subjects for training, validation, and testing which include 334, 84, and 140 scans respectively.

### 6.4.2 Methodology

Prediction of neonatal postmenstrual age at MRI using T2-weighted brain images consists of three parts. First, MRI images are segmented into different cortical and intra-cortical sub-structures using a state-of-the-art deep neural segmentation network. After that, features that are representative of neonatal brain age estimation are extracted for each structure separately. Once proper features for different substructures are extracted, a regression machine learning approach is used to predict the neonatal brain age. Each of these steps are detailed below.

#### 6.4.2.1 Brain segmentation

Convolutional neural networks (CNN) have proven to be a powerful tool for image segmentation(Minaee *et al.*, 2021). The use of CNNs can significantly improve the accuracy of brain segmentation which is a critical task in neuroimaging with various applications in both clinical and research settings (Fawzi, Achuthan & Belaton, 2021; Liu *et al.*, 2023b). CNNs can learn to automatically identify complex patterns and relationships within brain images, enabling precise and efficient segmentation of different brain structures. Motivated by recent advancements in deep learning segmentation models (Ronneberger *et al.*, 2015; Isensee *et al.*, 2021; Wang *et al.*, 2021c; Hatamizadeh *et al.*, 2022a), a CNN model was trained to segment the brain into 87 different regions by using the segmentation masks provided by DRAW-EM. The aim of the segmentation model is to assign a unique structure class to each voxel of the brain MRI so that there exists a maximum overlapping with respect to the ground-truth masks. This is achieved by optimizing the network parameters to minimize a segmentation loss function (e.g., Cross-Entropy or Dice loss (Milletari, Navab & Ahmadi, 2016)) for each pair of the input volume $\boldsymbol{x}_n$ and its corresponding segmentation volume $\boldsymbol{y}_n$ as:

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{seg} \left( \widetilde{\mathbf{y}}_n = f_{seg} \left( \boldsymbol{x}_n | \theta \right), \ \boldsymbol{y}_n \right). \tag{6.1}$$

Where $N$ is the number of training samples, $f_{seg}(.|\theta)$ is the segmentation network containing a set of learnable parameters $\theta$, $\mathcal{L}_{seg}$ is the segmentation loss function and $\widetilde{\mathbf{y}}_n$ the predicted segmentation.

#### 6.4.2.2 Extracting representative features

In the literature, different landmarks are reported to be important for neonatal brain age estimation. Nevertheless, they are either inaccurate or very hard and time-consuming to calculate (Schaer

*et al.*, 2012). In this work, we propose to use two different landmarks for segmented structures which are representative for brain age estimation.

The first type of features which are used to predict postmenstrual age from brain MRI is the relational volume (RV) of each structure compared to the whole brain. In particular, RV is computed by dividing the number of voxels assigned to a given structure by the total number of voxels contained in all the structures. The advantage of relational volume to absolute volume is that it removes the effect of head size for age prediction. Furthermore, it is very easy to compute, as basically involves summing the voxels across structures and a single division. As it can be seen in the top plot of Figure 6.3, there is a high correlation between the relational volume of several structures in the brain and postmenstrual age. This indicates that this metric has the potential to be a good indicator of brain age.

Another feature that is reported in the literature as a good indicator for brain age estimation is how gyrified and folded the brain cortex and other structures are. For example, neonates' brain cortex will have more gyrification as the preterm neonates grow. Nevertheless, computing the gyrification index and estimating the neonatal brain age based on this biomarker is very slow and hard to achieve (Schaer *et al.*, 2012). In this work, we propose using a novel measurement that is highly correlated with gyrification, but it is much easier to calculate. The ratio of the surface of a structure to its volume shows how folded a structure is. Thus, the surface to volume ratio (SVR) of a folded structure would be higher than the surface to volume ratio of a smooth and unfolded structure. Besides, this metric can be calculated easily by simply counting the number of voxels on the surface of a structure (i.e., neighboring voxels labels are different) and dividing it by the corresponding structure volume. We refer to this structure feature as SVR. Similarly, the plot in the bottom right of Figure 6.3 illustrates that there exists a strong correlation between the SVR of several regions and the postmenstrual age.

Based on these observations, the above-mentioned two features, i.e., relational volume (RV) and surface to volume ratio (SVR), were calculated for each of the 87 structures and they are later employed for training the brain age regression model.

#### 6.4.2.3 Postmenstrual age regression

To predict the neonatal brain age based on the extracted features from previous steps, we used a machine learning regression approach known as Bayesian Ridge (Tipping, 2001). The ridge regression model is defined as:

$$\arg\min_{\omega} \|Z - \chi\omega\|_2^2 + \lambda\|\omega\|_2^2 \tag{6.2}$$

where $Z$ represents the postmenstrual ages, $\chi$ are the extracted features from previous steps, $\omega$ the model parameters, and $\lambda$ a balancing term, which imposes a penalty on the size of the coefficients which makes the regression model more robust and generalized. Bayesian Ridge regression models linear regression using probability distribution rather than point estimates, which allows it to handle limited or poorly distributed data. To do so, the output $y$ is assumed to be Gaussian distributed around $\chi\omega$:

$$p(y \mid \chi, w, \alpha) = \mathcal{N}(y \mid \chi w, \alpha) \tag{6.3}$$

Also, Bayesian Ridge regression estimates a probabilistic model of the regression problem using spherical Gaussian prior for the coefficients.

$$p(w \mid \lambda) = \mathcal{N}\left(w \mid 0, \lambda^{-1}\mathbf{I}_p\right) \tag{6.4}$$

where the priors over $\alpha$ and $\lambda$ are chosen to be gamma distributions.

To improve the robustness of the model, we selected the strongest 100 correlated features with brain age and employed them to train the Bayesian ridge model.

For better comparison and analyzing the effect of the regression model, we also evaluated popular regression models including: Kernel Ridge regressor (Murphy, 2012), ElasticNet (Zou & Hastie, 2005), Gradient Boosting regressor (Friedman, 2002), Support Vector Machine (SVM) regressor (Drucker *et al.*, 1996), and Multi-layer Perceptron (MLP) regressor (Hinton, 1990).

#### 6.4.2.4 Discovering important correlations

Last, to investigate the importance of each structure and feature for brain age estimation, we utilized the PFI technique (Breiman, 2001). The permutation feature importance is defined as the drop in a trained model score caused by randomly shuffled feature values. In other words, it tries to capture the importance of each feature by measuring the decrease in accuracy when one feature is randomly permuted with other features. Because this approach removes the link between the feature and the goal, the decline in the model score reflects how much the model is dependent on the feature. This technique reveals the most important structures and features for brain age estimation which can be used as a reliable landmark for brain age.

### 6.4.3 Evaluation protocol

The accuracy of segmentation models was calculated using the average of Dice Coefficient Score (DSC) (Zou *et al.*, 2004) for all labels which measures the overlap between the generated segmentations and their corresponding ground-truth masks. For a given class, the DSC for a sample is formally defined as:

$$DSC = \frac{2|\hat{\mathbf{y}} \cdot \mathbf{y}|}{|\hat{\mathbf{y}}| + |\mathbf{y}|} \tag{6.5}$$

where $\hat{\mathbf{y}}$ is the discretized predicted segmentation, and $\mathbf{y}$ is the ground-truth segmentation.

Two metrics were used to assess the performance of the brain age estimation models: coefficient of determination ($R^2$) (Devore, 2011), which measures how well the predictions approximate the real ages, and Mean Absolute Error (MAE), which provides the average prediction error. For a set of $N$ samples, these metrics are formulated as following:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|z_i - \hat{z}_i| \tag{6.6}$$

$$R^2 = \left( \frac{N \sum_{i=1}^{N} z_i \hat{z}_i - \sum_{i=1}^{N} z_i \sum_{i=1}^{N} \hat{z}_i}{\sqrt{N \sum_{i=1}^{N} z_i^2 - \left( \sum_{i=1}^{N} z_i \right)^2} \sqrt{N \sum \hat{z}_i^2 - \left( \sum_{i=1}^{N} \hat{z}_i \right)^2}} \right)^2 \tag{6.7}$$

### 6.4.4 Compared methods

To evaluate the performance of the proposed method with respect to existing literature we included three recent relevant works in our empirical validation. These methods include: a model based on multiplanar slices and Test-Time Augmentation (2DTTA) (Hong *et al.*, 2021), an attention-based residual network (ARN) (Shi *et al.*, 2020) and a Global-Local Transformer (GLT) (He *et al.*, 2021). Note that these methods represent the state-of-the-art for image-based brain age estimation. Furthermore, it is noteworthy to mention that for all the methods, including 2DTTA, ARN, and GLT we run the experiments on the same data splits. Moreover, to have a fair comparison, we also searched for the best sets of hyperparameters and used the same data augmentation strategy for all tested methods.

### 6.4.5 Implementation details

All segmentation networks and regression methods were implemented using PyTorch. We trained the segmentation network using Adam optimizer with a learning rate starting at $1 \times 10^{-3}$, a weight decay of 0.5 every 20 epochs, and a batch-size of 32. Furthermore, these networks are trained on small 3D patches of sizes equal to $64 \times 64 \times 64$ voxels, following the standard literature in medical image segmentation. At test time, the final predicted segmentation is generated by sticking the segmentation of small patches together. Moreover, the regression models are based on the implemented models in scikit-learn library. Last, for all methods, we searched the optimal hyper-parameters on the independent validation set. Experiments were run in a server with 2 NVIDIA RTX A6000 GPU cards.

# CONCLUSION AND RECOMMENDATIONS

This thesis explored the use of generative models to advance brain MRI analysis, addressing several persistent challenges that hinder the clinical deployment of deep learning techniques. Specifically, we tackled the issues of inter-scanner variability, limited annotated data, and the subtlety of certain neurological conditions that elude conventional imaging analysis. Our contributions span three major domains: unsupervised MRI harmonization, unsupervised brain anomaly detection, and neonatal brain age estimation—each addressing a critical bottleneck in current neuroimaging pipelines.

First, we introduced **Harmonizing Flows, a source-free MRI harmonization** framework based on normalizing flows, which standardizes MRI appearance across sites without requiring paired or labeled data. This method demonstrated significant improvements in model generalizability and preservation of diagnostic content, making it suitable for large-scale, multi-site studies and real-world clinical deployment.

Second, we developed a series of generative models for **unsupervised brain anomaly detection**, culminating in **REFLECT**, a rectified flow-based method that achieves state-of-the-art localization and reconstruction of pathological regions. By training only on healthy data, our methods enable detection of rare or unseen anomalies, offering robust solutions in data-scarce settings and improving diagnostic support without reliance on exhaustive annotations.

Third, we proposed a deep learning framework for **neonatal brain age estimation** to assess developmental maturity directly from structural MRIs. By predicting brain age and identifying deviations from expected maturation, our model provides fine-grained, region-specific biomarkers that are especially valuable in detecting early-stage neurodevelopmental delays in neonates, a population where timely intervention is critical.

Together, these contributions form a cohesive generative-model-based framework for brain MRI assessment that is scalable, interpretable, and clinically relevant. They respond to key limitations in existing approaches and open new pathways for generalizable, label-efficient, and biologically grounded neuroimaging analysis.

**Recommendations for Future Research**

While the proposed methods show strong potential, several directions remain open for future investigation:

- **Unified generative frameworks:** Future work could explore integrating harmonization, anomaly detection, and brain age estimation into a single, multi-task architecture. This would enable shared representations and potentially enhance performance across tasks while reducing computational redundancy.
- **Multimodal integration:** Combining structural MRI with complementary modalities such as diffusion MRI, functional MRI, or clinical metadata could further improve the sensitivity and specificity of generative models, especially for complex or ambiguous cases.
- **Clinical translation and deployment:** Moving toward deployment requires robust uncertainty quantification, interpretability mechanisms, and integration with clinical workflows. Future work should focus on making generative models explainable and trustworthy for radiologists and clinicians.
- **Ethical and fairness considerations:** As generative models are increasingly used in clinical settings, it is essential to evaluate their fairness and ensure consistent performance across demographic groups, especially given the variability in healthcare access and imaging practices.

In closing, this thesis underscores the transformative potential of generative models in neuroimaging. By addressing key technical and clinical challenges, it contributes to the growing body of work aimed at building robust, scalable, and clinically impactful AI systems.

Continued research in this direction promises to reshape the landscape of medical imaging, ultimately leading to earlier diagnoses, more personalized treatments, and better outcomes for patients worldwide.

## BIBLIOGRAPHY


Abbasi, S., Lan, H., Choupan, J., Sheikh-Bahaei, N., Pandey, G. & Varghese, B. (2024). Deep learning for the harmonization of structural MRI scans: a survey. *BioMedical Engineering OnLine*, 23(1), 90.

Abdelhamed, A., Brubaker, M. A. & Brown, M. S. (2019). Noise flow: Noise modeling with conditional normalizing flows. *ICCV*, pp. 3165–3173.

Adamson, C. L., Alexander, B., Ball, G., Beare, R., Cheong, J. L., Spittle, A. J., Doyle, L. W., Anderson, P. J., Seal, M. L. & Thompson, D. K. (2020). Parcellation of the neonatal cortex using Surface-based Melbourne Children's Regional Infant Brain atlases (M-CRIB-S). *Scientific reports*, 10(1), 4359.

Aggarwal, H. K., Mani, M. P. & Jacob, M. (2018). MoDL: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2), 394–405.

Akrami, H., Joshi, A. A., Li, J., Aydore, S. & Leahy, R. M. (2020). Brain lesion detection using a robust variational autoencoder and transfer learning. *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pp. 786–790.

Alaverdyan, Z., Jung, J., Bouet, R. & Lartizien, C. (2020). Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: application to epilepsy lesion screening. *Medical image analysis*, 60, 101618.

Aldossary, N. M., Kotb, M. A. & Kamal, A. M. (2019). Predictive value of early MRI findings on neurocognitive and psychiatric outcomes in patients with severe traumatic brain injury. *Journal of affective disorders*, 243, 1–7.

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M. et al. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific data*, 4(1), 1–26.

Ali, S. S. A., Memon, K., Yahya, N. & Khan, S. (2025). Deep learning frameworks for MRI-based diagnosis of neurological disorders: a systematic review and meta-analysis. *Artificial Intelligence Review*, 58(6), 171.

Ashburner, J. & Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, 11(6), 805–821.

Atlason, H. E., Love, A., Sigurdsson, S., Gudnason, V. & Ellingsen, L. M. (2019). Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. *Medical Imaging 2019: Image Processing*, 10949, 372–378.

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K. & Davatzikos, C. (2017). Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1), 1–13.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M. et al. (2018). 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*.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. (2019). Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8), 1788–1800.

Basaia, S., Agosta, F., Wagner, L., Canu, E., Magnani, G., Santangelo, R., Filippi, M., Initiative, A. D. N. et al. (2019). Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. *NeuroImage: Clinical*, 21, 101645.

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H. & Ayed, I. B. (2022). Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 82, 102617.

Batzner, K., Heckler, L. & König, R. (2024). Efficientad: Accurate visual anomaly detection at millisecond-level latencies. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 128–138.

Baugh, M., Tan, J., Müller, J. P., Dombrowski, M., Batten, J. & Kainz, B. (2023). Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 162–172.

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2018). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. *International MICCAI brainlesion workshop*, pp. 161–169.

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2019). Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018*, pp. 161–169.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S. & Navab, N. (2020a). Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. *International conference on medical image computing and computer-assisted Intervention (MICCAI)*, pp. 718–727.

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2020b). Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pp. 1905–1909.

Baur, C., Wiestler, B., Albarqouni, S. & Navab, N. (2020c). Scale-space autoencoders for unsupervised anomaly segmentation in brain MRI. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pp. 552–561.

Baur, C., Denner, S., Wiestler, B., Navab, N. & Albarqouni, S. (2021a). Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Medical Image Analysis*, 69, 101952.

Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N. & Albarqouni, S. (2021b). Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain MRI. *Radiology: Artificial Intelligence*, 3(3), e190169.

Beer, J. C. et al. (2020). Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage*, 220, 117129.

Behrendt, F., Bengs, M., Rogge, F., Krüger, J., Opfer, R. & Schlaefer, A. (2022). Unsupervised anomaly detection in 3D brain MRI using deep learning with impured training data. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 1–4.

Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R. & Schlaefer, A. (2024a). Patched diffusion models for unsupervised anomaly detection in brain MRI. *Medical Imaging with Deep Learning*, pp. 1019–1032.

Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R. & Schlaefer, A. (2024b). Patched diffusion models for unsupervised anomaly detection in brain mri. *Medical Imaging with Deep Learning*, pp. 1019–1032.

Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R. & Schlaefer, A. (2024c). Leveraging the Mahalanobis Distance to Enhance Unsupervised Brain MRI Anomaly Detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 394–404.

Beizaee, F., Desrosiers, C., Lodygensky, G. A. & Dolz, J. (2023). Harmonizing Flows: Unsupervised MR harmonization based on normalizing flows. *International Conference on Information Processing in Medical Imaging*, pp. 347–359.

Beizaee, F., Lodygensky, G., Desrosiers, C. & Dolz, J. (2025). MAD-AD: Masked Diffusion for Unsupervised Brain Anomaly Detection. *International Conference on Information Processing in Medical Imaging*, pp. 139–153.

Bento, M., Fantini, I., Park, J., Rittner, L. & Frayne, R. (2022). Deep learning in large and multi-site structural brain MR imaging datasets. *Frontiers in Neuroinformatics*, 15, 805669.

Bercea, C. I., Neumayr, M., Rueckert, D. & Schnabel, J. A. (2023a). Mask, Stitch, and Re-Sample: Enhancing Robustness and Generalizability in Anomaly Detection through Automatic Diffusion Models. *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Bercea, C. I., Wiestler, B., Rueckert, D. & Schnabel, J. A. (2023b). Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 293–303.

Bercea, C. I., Wiestler, B., Rueckert, D. & Schnabel, J. A. (2024a). Diffusion models with implicit guidance for medical anomaly detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 211–220.

Bercea, C. I., Wiestler, B., Rueckert, D. & Schnabel, J. A. (2024b). Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. *Medical Imaging with Deep Learning*, pp. 39–52.

Bergmann, P., Fauser, M., Sattlegger, D. & Steger, C. (2019). MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. *PProceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9592–9600.

Bhatt, N., Prados, D. R., Hodzic, N., Karanassios, C. & Tizhoosh, H. (2021). Unsupervised detection of lung nodules in chest radiography using generative adversarial networks. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3842–3845.

Bitar, R., Leung, G., Perng, R., Tadros, S., Moody, A. R., Sarrazin, J., McGregor, C., Christakis, M., Symons, S., Nelson, A. et al. (2006). MR pulse sequences: what every radiologist wants to know but is afraid to ask. *Radiographics*, 26(2), 513–537.

Blencowe, H., Cousens, S., Chou, D., Oestergaard, M., Say, L., Moller, A.-B., Kinney, M., Lawn, J. & (see acknowledgement for full list), B. T. S. P. B. A. G. (2013). Born too soon: the global epidemiology of 15 million preterm births. *Reproductive health*, 10, 1–14.

Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. & de Bruijne, M. (2019). Semi-Supervised Medical Image Segmentation via Learning Consistency under Transformations. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, 11769(Lecture Notes in Computer Science), 810–818.

Boudiaf, M. et al. (2022). Parameter-free Online Test-time Adaptation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8344–8353.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Brinjikji, W., Lanzino, G., Kallmes, D. F. & Cloft, H. J. (2014). Cerebral aneurysm treatment is beginning to shift to low volume centers. *Journal of NeuroInterventional Surgery*, 6(5), 349–352.

Brown, R. W., Cheng, Y.-C. N., Haacke, E. M., Thompson, M. R. & Venkatesan, R. (2014). *Magnetic resonance imaging: physical principles and sequence design*. John Wiley & Sons.

Cackowski, S., Barbier, E. L., Dojat, M. & Christen, T. (2023). ImUnity: a generalizable VAE-GAN solution for multicenter MR image harmonization. *Medical Image Analysis*, 88, 102799.

Cai, Y., Chen, H., Yang, X., Zhou, Y. & Cheng, K.-T. (2023). Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical image analysis*, 86, 102794.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Klimstra, D. S. & Fuchs, T. J. (2019). Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole-Slide Images. *Nature Medicine*, 25(8), 1301–1309.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. *European conference on computer vision*, pp. 205–218.

Chalela, J. A., Kidwell, C. S., Nentwich, L. M., Luby, M., Butman, J. A., Demchuk, A. M., Hill, M. D., Patronas, N., Latour, L. & Warach, S. (2007). Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison. *The Lancet*, 369(9558), 293–298.

Chen, C., Dou, Q., Chen, H., Qin, J. & Heng, P. A. (2020a). Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7), 2494–2505.

Chen, J. V., Chaudhari, G., Hess, C. P., Glenn, O. A., Sugrue, L. P., Rauschecker, A. M. & Li, Y. (2022). Deep learning to predict neonatal and infant brain age from myelination on brain MRI scans. *Radiology*, 305(3), 678–687.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. (2018). Neural Ordinary Differential Equations. *NeurIPS*, pp. 6571–6583.

Chen, X. & Konukoglu, E. (2018). Unsupervised Detection of Lesions in Brain MRI using constrained adversarial auto-encoders. *Medical Imaging with Deep Learning*.

Chen, X., You, S., Tezcan, K. C. & Konukoglu, E. (2020b). Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64, 101713.

Cheng, J., Liu, Z., Guan, H., Wu, Z., Zhu, H., Jiang, J., Wen, W., Tao, D. & Liu, T. (2021). Brain age estimation from MRI using cascade networks with ranking loss. *IEEE Transactions on Medical Imaging*, 40(12), 3400–3412.

Cheong, J., Cameron, K. L. I., Thompson, D., Anderson, P. J., Ranganathan, S., Clark, R., Mentiplay, B., Burnett, A., Lee, K., Doyle, L. W. et al. (2021). Impact of moderate and late preterm birth on neurodevelopment, brain development and respiratory health at school age: protocol for a longitudinal cohort study (LaPrem study). *BMJ open*, 11(1), e044491.

Chiu, L.-L. & Lai, S.-H. (2023). Self-supervised normalizing flows for image anomaly detection and localization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2927–2936.

Cho, J., Kang, I. & Park, J. (2021). Self-supervised 3d out-of-distribution detection via pseudoanomaly generation. *International conference on medical image computing and computer-assisted intervention*, pp. 95–103.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S. & Vedaldi, A. (2014). Describing textures in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3606–3613.

Cohen, J. P., Luck, M. & Honari, S. (2018). Distribution matching losses can hallucinate features in medical image translation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 529–536.

Cole, J. H. e. a. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163, 115–124.

Croitoru, F.-A., Hondru, V., Ionescu, R. T. & Shah, M. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 10850–10869.

Dale, B. M., Brown, M. A. & Semelka, R. C. (2015). *MRI: basic principles and applications*. John Wiley & Sons.

Defard, T., Setkov, A., Loesch, A. & Audigier, R. (2021). Padim: a patch distribution modeling framework for anomaly detection and localization. *International Conference on Pattern Recognition*, pp. 475–489.

Delisle, P.-L. et al. (2021). Realistic image normalization for multi-Domain segmentation. *Medical Image Analysis*, 74, 102191.

Deng, H. & Li, X. (2022). Anomaly detection via reverse distillation from one-class embedding. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9737–9746.

Deprez, M., Wang, S., Ledig, C., Hajnal, J. V., Counsell, S. J. & Schnabel, J. A. (2018). Segmentation of myelin-like signals on clinical MR images for age estimation in preterm infants. *bioRxiv*, 357749.

Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences*. Cengage learning.

Dewey, B. E. et al. (2019). DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magnetic resonance imaging*, 64, 160–170.

Dewey, B. E. et al. (2020). A disentangled latent space for cross-site MRI harmonization. *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 720–729.

Di Martino, A. et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6), 659–667.

Dinh, L., Krueger, D. & Bengio, Y. (2015). Nice: Non-linear independent components estimation. *Workshop paper at International Conference on Learning Representations*.

Dinh, L. et al. (2017). Density estimation using real NVP. *International Conference on Learning Representations (ICLR)*.

Dinsdale, N. K. et al. (2021). Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage*, 228, 117689.

Dolz, J., Desrosiers, C. & Ayed, I. B. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170, 456–470.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X. & Heng, P.-A. (2019). Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7, 99065–99076.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. & Vapnik, V. (1996). Support vector regression machines. *Advances in neural information processing systems*, 9.

Dubois, J., Benders, M., Cachia, A., Lazeyras, F., Ha-Vinh Leuchter, R., Sizonenko, S. V., Borradori-Tolsa, C., Mangin, J.-F. & Hüppi, P. S. (2008). Mapping the early cortical folding process in the preterm newborn brain. *Cerebral cortex*, 18(6), 1444–1454.

Dubois, J., Dehaene-Lambertz, G., Kulikova, S., Poupon, C., Hüppi, P. S. & Hertz-Pannier, L. (2014). The early development of brain white matter: a review of imaging studies in fetuses, newborns and infants. *neuroscience*, 276, 48–71.

Durairaj, A., Madhan, E., Rajkumar, M. & Shameem, S. (2024). Optimizing anomaly detection in 3D MRI scans: The role of ConvLSTM in medical image analysis. *Applied Soft Computing*, 164, 111919.

Durrer, A., Wolleb, J., Bieder, F., Sinnecker, T., Weigel, M., Sandkuehler, R., Granziera, C., Yaldizli, Ö. & Cattin, P. C. (2023). Diffusion Models for Contrast Harmonization of Magnetic Resonance Images. *Medical Imaging with Deep Learning*.

Ellingson, B. M., Bendszus, M., Sorensen, A. G. & Pope, W. B. (2014). Emerging techniques and technologies in brain tumor imaging. *Neuro-oncology*, 16(suppl_7), vii12–vii23.

Engelhardt, E., Inder, T. E., Alexopoulos, D., Dierker, D. L., Hill, J., Van Essen, D. & Neil, J. J. (2015). Regional impairments of cortical folding in premature infants. *Annals of neurology*, 77(1), 154–162.

Esser, P., Rombach, R. & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12873–12883.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F. et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. *ICML*.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine*, 25(1), 24–29.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J. & Socher, R. (2021). Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1), 1–9.

Fawzi, A., Achuthan, A. & Belaton, B. (2021). Brain image segmentation in recent years: A narrative review. *Brain Sciences*, 11(8), 1055.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62(2), 774–781.

Flaborea, A., Collorone, L., Di Melendugno, G. M. D., D'Arrigo, S., Prenkaj, B. & Galasso, F. (2023). Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10318–10329.

Fleiss, B., Wong, F., Brownfoot, F., Shearer, I. K., Baud, O., Walker, D. W., Gressens, P. & Tolcos, M. (2019). Knowledge gaps and emerging research areas in intrauterine growth restriction-associated brain injury. *Frontiers in endocrinology*, 10, 188.

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., Adams, P., Cooper, C., Fava, M., McGrath, P. J. et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*, 167, 104–120.

Fortin, J.-P. et al. (2016). Removing inter-subject technical variability in magnetic resonance imaging studies. *NeuroImage*, 132, 198–212.

Fortin, J.-P. et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161, 149–170.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), 367–378.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200), 675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The annals of mathematical statistics*, 11(1), 86–92.

Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P. & Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nature reviews neurology*, 6(2), 67–77.

Fučka, M., Zavrtanik, V. & Skočaj, D. (2024). TransFusion–A transparency-based diffusion model for anomaly detection. *European conference on computer vision*, pp. 91–108.

Galdi, P., Blesa, M., Stoye, D. Q., Sullivan, G., Lamb, G. J., Quigley, A. J., Thrippleton, M. J., Bastin, M. E. & Boardman, J. P. (2020). Neonatal morphometric similarity mapping for predicting brain age and characterizing neuroanatomic variation associated with preterm birth. *NeuroImage: Clinical*, 25, 102195.

Gao, Y., Liu, Y., Wang, Y., Shi, Z. & Yu, J. (2019). A universal intensity standardization method based on a many-to-one weak-paired cycle generative adversarial network for magnetic resonance images. *IEEE transactions on medical imaging*, 38(9), 2059–2069.

Gholipour, A. e. a. (2017). A review of structural and functional brain imaging in preterm infants. *NeuroImage: Clinical*, 15, 296–310.

Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *arXiv preprint arXiv:1910.04597*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative Adversarial Networks. *NeurIPS*, pp. 2672–2680.

Graham, M. S., Pinaya, W. H., Tudosiu, P.-D., Nachev, P., Ourselin, S. & Cardoso, J. (2023). Denoising diffusion models for out-of-distribution detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2948–2957.

Grover, A. et al. (2020). Alignflow: Cycle consistent learning from multiple domains via normalizing flows. *AAAI*, pp. 4028–4035.

Guan, H., Liu, Y., Yang, E., Yap, P.-T., Shen, D. & Liu, M. (2021). Multi-site MRI harmonization via attention-guided deep domain adaptation for brain disorder identification. *Medical image analysis*, 71, 102076.

Guan, H., Liu, S., Lin, W., Yap, P.-T. & Liu, M. (2022). Fast image-level MRI harmonization via spectrum analysis. *International Workshop on Machine Learning in Medical Imaging*, pp. 201–209.

Gudovskiy, D. et al. (2022). Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. *WACV*, pp. 98–107.

Gui, L., Loukas, S., Lazeyras, F., Hüppi, P. S., Meskaldji, D. E. & Tolsa, C. B. (2019). Longitudinal study of neonatal brain tissue volumes in preterm infants and their ability to predict neurodevelopmental outcome. *Neuroimage*, 185, 728–741.

Guo, J., Lu, S., Jia, L., Zhang, W. & Li, H. (2023a). Encoder-decoder contrast for unsupervised anomaly detection in medical images. *IEEE transactions on medical imaging*, 43(3), 1102–1112.

Guo, J., Lu, S., Jia, L., Zhang, W. & Li, H. (2023b). Recontrast: Domain-specific anomaly detection via contrastive reconstruction. *Advances in Neural Information Processing Systems*, 36, 10721–10740.

Gutta, S., Acharya, J., Shiroishi, M., Hwang, D. & Nayak, K. S. (2021). Improved glioma grading using deep convolutional neural networks. *American Journal of Neuroradiology*, 42(2), 233–239.

Han, C., Rundo, L., Murao, K., Noguchi, T., Shimahara, Y., Milacski, Z. Á., Koshino, S., Sala, E., Nakayama, H. & Satoh, S. (2021). MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. *BMC bioinformatics*, 22, 1–20.

Han, Y., Sunwoo, L. & Ye, J. C. (2019). k-space deep learning for accelerated MRI. *IEEE transactions on medical imaging*, 39(2), 377–386.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R. & Xu, D. (2022a). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R. & Xu, D. (2022b). Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M. & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35, 18–31.

He, H., Bai, Y., Zhang, J., He, Q., Chen, H., Gan, Z., Wang, C., Li, X., Tian, G. & Xie, L. (2024a). MambaAD: Exploring State Space Models for Multi-class Unsupervised Anomaly Detection. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C. & Xie, L. (2024b). A diffusion-based framework for multi-class anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8), 8472–8480.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16000–16009.

He, S., Grant, P. E. & Ou, Y. (2021). Global-Local transformer for brain age estimation. *IEEE Transactions on Medical Imaging*, 41(1), 213–224.

Heeger, D. J. & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature reviews neuroscience*, 3(2), 142–151.

Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J. L., Sugano, N. & Sato, Y. (2018). Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size. *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pp. 31–41.

Hill, J., Dierker, D., Neil, J., Inder, T., Knutsen, A., Harwell, J., Coalson, T. & Van Essen, D. (2010a). A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. *Journal of neuroscience*, 30(6), 2268–2276.

Hill, J., Inder, T., Neil, J., Dierker, D., Harwell, J. & Van Essen, D. (2010b). Similar patterns of cortical expansion during human development and evolution. *Proceedings of the National Academy of Sciences*, 107(29), 13135–13140.

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine learning* (pp. 555–610). Elsevier.

Hintz, S. R. (2015). Neuroimaging and neurodevelopmental outcome in preterm infants. *Seminars in Perinatology*, 39(2), 132–140.

Ho, J., Jain, A. & Abbeel, P. (2020a). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840–6851.

Ho, J., Jain, A. & Abbeel, P. (2020b). Denoising diffusion probabilistic models. *NeurIPS*, 33, 6840–6851.

Ho, J. et al. (2019). Flow++: Improving flow-based generative models with variational dequantization and architecture design. *ICML*, pp. 2722–2730.

Hong, J., Yun, H. J., Park, G., Kim, S., Ou, Y., Vasung, L., Rollins, C. K., Ortinau, C. M., Takeoka, E., Akiyama, S. et al. (2021). Optimal method for fetal brain age prediction using multiplanar slices from structural magnetic resonance imaging. *Frontiers in neuroscience*, 1284.

Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y. & Zhang, S. (2021). Fully test-time adaptation for image segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pp. 251–260.

Hughes, E., Cordero-Grande, L., Murgasova, M., Hutter, J., Price, A., Gomes, A. D. S., Allsop, J., Steinweg, J., Tusor, N., Wurie, J. et al. (2017). The Developing Human Connectome: announcing the first release of open access neonatal brain imaging. *Organization for Human Brain Mapp*, 25–29.

Hüppi, P. S., Warfield, S., Kikinis, R., Barnes, P. D., Zientara, G. P., Jolesz, F. A., Tsuji, M. K. & Volpe, J. J. (1998). Quantitative magnetic resonance imaging of brain development in premature and mature newborns. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 43(2), 224–235.

Iqbal, H., Khalid, U., Chen, C. & Hua, J. (2023). Unsupervised anomaly detection in medical images using masked diffusion model. *International Workshop on Machine Learning in Medical Imaging*, pp. 372–381.

Isensee, F. et al. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2), 203–211.

IXI Dataset. [Accessed April 2025]. (2004). IXI Dataset - Brain Development. Retrieved from: https://brain-development.org/ixi-dataset/.

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W. & Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1), 119–128.

Jeong, H., Byun, H., Kang, D. U. & Lee, J. (2023). BlindHarmony:" Blind" Harmonization for MR Images via Flow model. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21129–21139.

Jeong, H., Lee, H., Chun, S. Y. & Lee, J. (2025). Efficient and robust 3D blind harmonization for large domain gaps. *arXiv preprint arXiv:2505.00133*.

Jezek, S., Jonak, M., Burget, R., Dvorak, P. & Skotak, M. (2021). Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pp. 66–71.

Johnson, W. E., Li, C. & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 118–127.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D. et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pp. 597–609.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D. et al. (2018). Ensembles of multiple models and architectures for robust brain tumour segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*, pp. 450–462.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management science*, 6(4), 366–422.

Karani, N., Chaitanya, K., Baumgartner, C. & Konukoglu, E. (2018). A lifelong learning approach to brain MR segmentation across scanners and protocols. *International conference on medical image computing and computer-assisted Intervention (MICCAI)*, pp. 476–484.

Karani, N. et al. (2021). Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68, 101907.

Kascenas, A., Pugeault, N. & O'Neil, A. Q. (2022). Denoising autoencoders for unsupervised anomaly detection in brain MRI. *International Conference on Medical Imaging with Deep Learning*, pp. 653–664.

Kelly, C., Ball, G., Matthews, L. G., Cheong, J. L., Doyle, L. W., Inder, T. E., Thompson, D. K. & Anderson, P. J. (2022). Investigating brain structural maturation in children and adolescents born very preterm using the brain age framework. *NeuroImage*, 247, 118828.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J. & Ben, I. (2021). Boundary loss for highly unbalanced segmentation. *Medical Image Analysis*, 67(10185), 101851.

Kim, D., Baik, S. & Kim, T. H. (2023). Sanflow: Semantic-aware normalizing flow for anomaly detection. *NeurIPS*, 36, 75434–75454.

Kingma, D., Salimans, T., Poole, B. & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, 34, 21696–21707.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. & Dhariwal, P. (2018). Glow: generative flow with invertible 1× 1 convolutions. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10236–10245.

Kingma, D. P. & Welling, M. (2014). Auto-Encoding Variational Bayes. *stat*, 1050, 1.

Kirichenko, P., Izmailov, P. & Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. *NeurIPS*, 33, 20578–20589.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., Fox, N. C., Jack Jr, C. R., Ashburner, J. & Frackowiak, R. S. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain*, 131(3), 681–689.

Korolev, S., Safiullin, A., Belyaev, M. & Dodonova, Y. (2017). Residual and plain convolutional neural networks for 3D brain MRI classification. *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pp. 835–838.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kuijf, H. J., Biesbroek, J. M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M. J., Casamitjana, A. et al. (2019). Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge. *IEEE transactions on medical imaging*, 38(11), 2556–2568.

Kumar, M., Beyea, S., Hu, S. & Kamal, N. (2024). Impact of early MRI in ischemic strokes beyond hyper-acute stage to improve patient outcomes, enable early discharge, and realize cost savings. *Journal of Stroke and Cerebrovascular Diseases*, 33(5), 107662.

Landfeldt, E., Castelo-Branco, A., Svedbom, A., Löfroth, E., Kavaliunas, A. & Hillert, J. (2018). The long-term impact of early treatment of multiple sclerosis on the risk of disability pension. *Journal of Neurology*, 265, 701–707.

Le Bihan, D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature reviews neuroscience*, 4(6), 469–480.

Lee, H., Yang, Y., Xu, J., Ware, J. B. & Liu, B. (2021). Use of magnetic resonance imaging in acute traumatic brain injury patients is associated with lower inpatient mortality. *Journal of clinical imaging science*, 11, 53.

Lehéricy, S., Sharman, M. A., Dos Santos, C. L., Paquin, R. & Gallea, C. (2012). Magnetic resonance imaging of the substantia nigra in Parkinson's disease. *Movement Disorders*, 27(7), 822–830.

Li, C.-L., Sohn, K., Yoon, J. & Pfister, T. (2021). Cutpaste: Self-supervised learning for anomaly detection and localization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9664–9674.

Li, H., Chen, Z., Xu, Y. & Hu, J. (2024). Hyperbolic Anomaly Detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 17511–17520.

Liang, J. et al. (2020). Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. *ICML*, pp. 6028–6039.

Liang, Z., Anthony, H., Wagner, F. & Kamnitsas, K. (2023). Modality cycles with masked conditional diffusion for unsupervised anomaly segmentation in MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 168–181.

Liang, Z., Guo, X., Noble, J. A. & Kamnitsas, K. (2024a). IterMask 2: Iterative Unsupervised Anomaly Segmentation via Spatial and Frequency Masking for Brain Lesions in MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 339–348.

Liang, Z., Guo, X., Noble, J. A. & Kamnitsas, K. (2024b). IterMask 2: Iterative Unsupervised Anomaly Segmentation via Spatial and Frequency Masking for Brain Lesions in MRI. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 339–348.

Liew, S.-L., Lo, B. P., Donnelly, M. R., Zavaliangos-Petropulu, A., Jeong, J. N., Barisano, G., Hutton, A., Simon, J. P., Juliano, J. M., Suri, A. et al. (2022). A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1), 320.

Lin, W. (2000). Principles of magnetic resonance imaging: a signal processing perspective [Book Review]. *IEEE Engineering in Medicine and Biology Magazine*, 19(5), 129–130.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. (2022). Flow Matching for Generative Modeling. *The Eleventh International Conference on Learning Representations*.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42, 60–88.

Liu, L., Oza, S., Hogan, D., Chu, Y., Perin, J., Zhu, J., Lawn, J. E., Cousens, S., Mathers, C. & Black, R. E. (2016). Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*, 388(10063), 3027–3035.

Liu, M., Lu, M., Kim, S. Y., Lee, H. J., Duffy, B. A., Yuan, S., Chai, Y., Cole, J. H., Wu, X., Toga, A. W. et al. (2024). Brain age predicted using graph convolutional neural network explains neurodevelopmental trajectory in preterm neonates. *European Radiology*, 34(6), 3601–3611.

Liu, M. et al. (2021). Style transfer using generative adversarial networks for multi-site MRI harmonization. *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 313–322.

Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., Bhanu, B., Radke, R. J. & Camps, O. (2020). Towards visually explaining variational autoencoders. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8642–8651.

Liu, X., Gong, C. & Liu, Q. (2023a). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *The Eleventh International Conference on Learning Representations (ICLR)*.

Liu, Z., Tong, L., Chen, L., Jiang, Z., Zhou, F., Zhang, Q., Zhang, X., Jin, Y. & Zhou, H. (2023b). Deep learning based brain tumor segmentation: a survey. *Complex & intelligent systems*, 9(1), 1001–1026.

Liu, Z., Zhou, Y., Xu, Y. & Wang, Z. (2023c). Simplenet: A simple network for image anomaly detection and localization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 20402–20411.

Lu, F., Yao, X., Fu, C.-W. & Jia, J. (2023). Removing anomalies as noises for industrial defect localization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16166–16175.

Lu, S., Zhang, W., Guo, J., Liu, H., Li, H. & Wang, N. (2024). PatchCL-AE: Anomaly detection for medical images using patch-wise contrastive learning-based auto-encoder. *Computerized Medical Imaging and Graphics*, 114, 102366.

Lucas, J., Tucker, G., Grosse, R. & Norouzi, M. (2019). Understanding posterior collapse in generative latent variable models. *DGS workshop at International Conference on Learning Representations (ICLR)*.

Lüth, C. T., Zimmerer, D., Koehler, G., Jaeger, P. F., Isenensee, F. & Maier-Hein, K. H. (2023). Contrastive representations for unsupervised anomaly detection and localization. *BVM Workshop*, pp. 246–252.

Ma, Y., Wang, D., Liu, P., Masters, L., Barnett, M., Cai, W. & Wang, C. (2024). Symmetry awareness encoded deep learning framework for brain imaging analysis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 742–752.

Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., Edwards, A. D., Counsell, S. J. & Rueckert, D. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE transactions on medical imaging*, 33(9), 1818–1831.

Makropoulos, A., Aljabar, P., Wright, R., Hüning, B., Merchant, N., Arichi, T., Tusor, N., Hajnal, J. V., Edwards, A. D., Counsell, S. J. et al. (2016). Regional growth and atlasing of the developing human brain. *Neuroimage*, 125, 456–478.

Makropoulos, A., Robinson, E. C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S. J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J. et al. (2018). The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage*, 173, 88–112.

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C. & Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9), 1498–1507.

Marimont, S. N., Baugh, M., Siomos, V., Tzelepis, C., Kainz, B. & Tarroni, G. (2024). Disyre: Diffusion-inspired synthetic restoration for unsupervised anomaly detection. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5.

Martins, S. B., Benato, B. C., Silva, B. F., Yasuda, C. L. & Falcão, A. X. (2019). Modeling normal brain asymmetry in MR images applied to anomaly detection without segmentation and data annotation. *Medical Imaging 2019: Computer-Aided Diagnosis*, 10950, 71–80.

Mathur, A. M., Neil, J. J., McKinstry, R. C. & Inder, T. E. (2008). Transport, monitoring, and successful brain MR imaging in unsedated neonates. *Pediatric radiology*, 38(3), 260–264.

Meissen, F., Kaissis, G. & Rueckert, D. (2021). Autoseg-steering the inductive biases for automatic pathology segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 127–135.

Meng, S., Meng, W., Zhou, Q., Li, S., Hou, W. & He, S. (2024). MoEAD: A Parameter-efficient Model for Multi-class Anomaly Detection. *European Conference on Computer Vision*.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R. et al. (2014). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10), 1993–2024.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7), 3523–3542.

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C. & Foresti, G. L. (2021). VT-ADL: A vision transformer network for image anomaly detection and localization. *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pp. 01–06.

Modanwal, G., Vellal, A., Buda, M. & Mazurowski, M. A. (2020). MRI image harmonization using cycle-consistent generative adversarial network. *Medical Imaging 2020: Computer-Aided Diagnosis*, 11314, 259–264.

Mummadi, C. K. et al. (2022). Test-time adaptation to distribution shift by confidence maximization and input transformation.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Murugesan, B., Vasudeva, S. A., Liu, B., Lombaert, H., Ayed, I. B. & Dolz, J. (2025). Neighbor-aware calibration of segmentation networks with penalty-based constraints. *Medical Image Analysis*, 103501.

Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B. & Tarroni, G. (2024). Ensembled Cold-Diffusion Restorations for Unsupervised Anomaly Detection. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 243–253.

Nguyen, B., Feldman, A., Bethapudi, S., Jennings, A. & Willcocks, C. G. (2021). Unsupervised region-based anomaly detection in brain MRI with adversarial image inpainting. *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1127–1131.

Nichol, A. Q. & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. *International conference on machine learning*, pp. 8162–8171.

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P. & Tan, M. (2022a). Efficient test-time model adaptation without forgetting. *International conference on machine learning*, pp. 16888–16905.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P. & Tan, M. (2022b). Towards Stable Test-time Adaptation in Dynamic Wild World. *The Eleventh International Conference on Learning Representations (ICLR)*.

Nyúl, L. G. & Udupa, J. K. (1999). On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6), 1072–1081.

Nyúl, L. G., Udupa, J. K. & Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE transactions on medical imaging*, 19(2), 143–150.

Osowiechi, D., Hakim, G. A. V., Noori, M., Cheraghalikhani, M., Ben Ayed, I. & Desrosiers, C. (2023). Tttflow: Unsupervised test-time training with normalizing flow. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2126–2134.

Pagnozzi, A. M., Fripp, J. & Rose, S. E. (2019). Quantifying deep grey matter atrophy using automated segmentation approaches: A systematic review of structural MRI studies. *Neuroimage*, 201, 116018.

Pang, G., Yan, C., Shen, C., Hengel, A. v. d. & Bai, X. (2020). Self-trained deep ordinal regression for end-to-end video anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12173–12182.

Parida, A., Jiang, Z., Packer, R. J., Avery, R. A., Anwar, S. M. & Linguraru, M. G. (2024). Quantitative Metrics for Benchmarking Medical Image Harmonization. *IEEE International Symposium on Biomedical Imaging (ISBI)*.

Peebles, W. & Xie, S. (2023). Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205.

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A. & Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68, 101871.

Perera, P., Nallapati, R. & Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2898–2906.

Pham, C.-H., Tor-Díez, C., Meunier, H., Bednarek, N., Fablet, R., Passat, N. & Rousseau, F. (2019). Multiscale brain MRI super-resolution using deep 3D convolutional networks. *Computerized Medical Imaging and Graphics*, 77, 101647.

Pinaya, W. H., Graham, M. S., Gray, R., Da Costa, P. F., Tudosiu, P.-D., Wright, P., Mah, Y. H., MacKinnon, A. D., Teo, J. T., Jager, R. et al. (2022a). Fast unsupervised brain anomaly detection and segmentation with diffusion models. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 705–714.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S. & Cardoso, M. J. (2022b). Brain imaging generation with latent diffusion models. *MICCAI Workshop on Deep Generative Models*, pp. 117–126.

Pinon, N., Trombetta, R. & Lartizien, C. (2024). One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities. *Medical Imaging with Deep Learning*, pp. 1783–1797.

Polman, C. H., Reingold, S. C., Banwell, B., Clanet, M., Cohen, J. A., Filippi, M. & Wolinsky, J. S. (2011). Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of Neurology*, 69(2), 292–302.

Pomponio, R. et al. (2020). Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208, 116450.

Pope, W. B., Sayre, J., Perlina, A., Villablanca, J. P., Mischel, P. S. & Cloughesy, T. F. (2005). MR imaging correlates of survival in patients with high-grade gliomas. *American Journal of Neuroradiology*, 26(10), 2466–2474.

Rezende, D. J. & Mohamed, S. (2015). Variational inference with normalizing flows. *ICML*, pp. 1530–1538.

Roca, V., Kuchcinski, G., Pruvo, J.-P., Manouvriez, D., Lopes, R. et al. (2025). IGUANe: A 3D generalizable CycleGAN for multicenter harmonization of brain MR images. *Medical Image Analysis*, 99, 103388.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10684–10695.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T. & Gehler, P. (2022). Towards total recall in industrial anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14318–14328.

Rudolph, M., Wehrbein, T., Rosenhahn, B. & Wandt, B. (2022). Fully convolutional cross-scale-flows for image-based defect detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1088–1097.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E. & Kloft, M. (2018). Deep one-class classification. *International conference on machine learning*, pp. 4393–4402.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. In Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1* (pp. 318–362). Cambridge, MA, USA: MIT Press.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H. & Rabiee, H. R. (2021). Multiresolution knowledge distillation for anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14902–14912.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X. (2016). Improved techniques for training GANs. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2234–2242.

Sato, K., Hama, K., Matsubara, T. & Uehara, K. (2019). Predictable uncertainty-aware unsupervised deep anomaly segmentation. *2019 international joint conference on neural networks (ijcnn)*, pp. 1–7.

Schaer, M., Cuadra, M. B., Schmansky, N., Fischl, B., Thiran, J.-P. & Eliez, S. (2012). How to measure cortical folding from MR images: a step-by-step tutorial to compute local gyrification index. *JoVE (Journal of Visualized Experiments)*, (59), e3417.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *International conference on information processing in medical imaging*, pp. 146–157.

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G. & Schmidt-Erfurth, U. (2019a). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. (2019b). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54, 30–44.

Schlüter, H. M., Tan, J., Hou, B. & Kainz, B. (2022). Natural synthetic anomalies for self-supervised anomaly detection and localization. *European Conference on Computer Vision*, pp. 474–489.

Schwarz, A. J. (2021). The use, standardization, and interpretation of brain imaging data in clinical trials of neurodegenerative disorders. *Neurotherapeutics*, 18(2), 686–708.

Shen, J. & Shen, H.-W. (2023). PSRFlow: Probabilistic super resolution with flow-based models for scientific data. *IEEE Transactions on Visualization and Computer Graphics*, 30(1), 986–996.

Shen, L., Zheng, J., Lee, E. H., Shpanskaya, K., McKenna, E. S., Atluri, M. G., Plasto, D., Mitchell, C., Lai, L. M., Guimaraes, C. V. et al. (2022). Attention-guided deep learning for gestational age prediction using fetal brain MRI. *Scientific reports*, 12(1), 1408.

Shenton, M. E., Hamoda, H. M., Schneiderman, J. S., Bouix, S., Pasternak, O., Rathi, Y., Vu, M.-A., Purohit, M. P., Helmer, K., Koerte, I. et al. (2012). A review of magnetic resonance imaging and diffusion tensor imaging findings in mild traumatic brain injury. *Brain imaging and behavior*, 6, 137–192.

Shi, W., Yan, G., Li, Y., Li, H., Liu, T., Sun, C., Wang, G., Zhang, Y., Zou, Y. & Wu, D. (2020). Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. *Neuroimage*, 223, 117316.

Shimony, J. S., Smyser, C. D., Wideman, G., Alexopoulos, D., Hill, J., Harwell, J., Dierker, D., Van Essen, D. C., Inder, T. E. & Neil, J. J. (2016). Comparison of cortical folding measures for evaluation of developing human brain. *Neuroimage*, 125, 780–790.

Shinohara, R. et al. (2014). Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6, 9–19.

Silva-Rodríguez, J., Naranjo, V. & Dolz, J. (2022). Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80, 102526.

Sinha, A. & Dolz, J. (2020). Multi-scale self-guided attention for medical image segmentation. *IEEE journal of biomedical and health informatics*, 25(1), 121–130.

Smith, D. H., Hicks, R. & Povlishock, J. T. (2013). Therapy development for diffuse axonal injury. *Journal of neurotrauma*, 30(5), 307–323.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E. et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23, S208–S219.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, pp. 2256–2265.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A. & Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33, 596–608.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J. & Goldstein, T. (2023). Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36, 47783–47803.

Song, J., Meng, C. & Ermon, S. (2021a). Denoising Diffusion Implicit Models. *International Conference on Learning Representations*.

Song, Y., Durkan, C., Murray, I. & Ermon, S. (2021b). Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34, 1415–1428.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. (2021c). Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations*.

Song, Y., Shen, L., Xing, L. & Ermon, S. (2022). s. *International Conference on Learning Representations*.

Spittle, A. J., Thompson, D. K., Brown, N. C., Treyvaud, K., Cheong, J. L., Lee, K. J., Pace, C. C., Olsen, J., Allinson, L. G., Morgan, A. T. et al. (2014). Neurobehaviour between birth and 40 weeks' gestation in infants born< 30 weeks' gestation and parental psychological wellbeing: predictors of brain development and child outcomes. *BMC pediatrics*, 14, 1–13.

Stan, S. & Rostami, M. (2024). Unsupervised model adaptation for source-free segmentation of medical images. *Medical Image Analysis*, 95, 103179.

Sun, H., Mehta, S., Khaitova, M., Cheng, B., Hao, X., Spann, M. & Scheinost, D. (2024). Brain age prediction and deviations from normative trajectories in the neonatal connectome. *Nature Communications*, 15(1), 10251.

Tabak, E. G. & Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2), 145–164.

Takao, H. et al. (2011). Effect of scanner in longitudinal studies of brain volume changes. *Journal of Magnetic Resonance Imaging*, 34(2), 438–444.

Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D. & Kainz, B. (2021). Detecting outliers with poisson image interpolation. *MICCAI*, pp. 581–591.

Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B. et al. (2022). Detecting Outliers with Foreign Patch Interpolation. *Machine Learning for Biomedical Imaging*, 1(April 2022 issue), 1–27.

Tanenbaum, L. N., Tsiouris, A. J., Johnson, A. N., Naidich, T. P., DeLano, M. C., Melhem, E. R., Quarterman, P., Parameswaran, S., Shankaranarayanan, A., Goyen, M. et al. (2017). Synthetic MRI for clinical neuroimaging: results of the magnetic resonance image compilation (MAGiC) prospective, multicenter, multireader trial. *American journal of neuroradiology*, 38(6), 1103–1110.

Taoudi-Benchekroun, Y., Christiaens, D., Grigorescu, I., Gale-Grant, O., Schuh, A., Pietsch, M., Chew, A., Harper, N., Falconer, S., Poppe, T. et al. (2022). Predicting age and clinical risk from the neonatal connectome. *NeuroImage*, 119319.

Tarvainen, A. & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.

Tian, D., Zeng, Z., Sun, X., Tong, Q., Li, H., He, H., Gao, J.-H., He, Y. & Xia, M. (2022). A deep learning-based multisite neuroimage harmonization framework established with a traveling-subject dataset. *NeuroImage*, 257, 119297.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun), 211–244.

Torbati, M. E., Tudorascu, D. L., Minhas, D. S., Maillard, P., DeCarli, C. S. & Hwang, S. J. (2021). Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3284–3293.

Tur, A. O., Dall'Asen, N., Beyan, C. & Ricci, E. (2023). Exploring diffusion models for unsupervised video anomaly detection. *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2540–2544.

Turner, M. R., Agosta, F., Bede, P., Govind, V., Lulé, D. & Verstraete, E. (2012). Neuroimaging in amyotrophic lateral sclerosis. *Biomarkers in Medicine*, 6(3), 319–337.

Usman, B. et al. (2020). Log-likelihood ratio minimizing flows: Towards robust and quantifiable neural distribution alignment. *NeurIPS*, 33, 21118–21129.

Uzunova, H., Schultz, S., Handels, H. & Ehrhardt, J. (2019). Unsupervised pathology detection in medical images using conditional variational autoencoders. *International journal of computer assisted radiology and surgery*, 14, 451–461.

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*, pp. 36–46.

van der Graaf, M. (2010). In vivo magnetic resonance spectroscopy: basic methodology and clinical applications. *European Biophysics Journal*, 39(4), 527–540.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, C., Zhu, W., Gao, B.-B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M. & Ma, L. (2024). Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 22883–22892.

Wang, D. et al. (2021a). TENT: Fully Test-Time Adaptation by Entropy Minimization. *International Conference on Learning Representations (ICLR)*.

Wang, G. et al. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34–45.

Wang, J., Levman, J., Pinaya, W. H. L., Tudosiu, P.-D., Cardoso, M. J. & Marinescu, R. (2023a). Inversesr: 3d brain mri super-resolution using a latent diffusion model. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 438–447.

Wang, P., Peng, J., Pedersoli, M., Zhou, Y., Zhang, C. & Desrosiers, C. (2023b). Shape-aware joint distribution alignment for cross-domain image segmentation. *IEEE Transactions on Medical Imaging*, 42(8), 2338–2347.

Wang, R. et al. (2021b). Harmonization with flow-based causal inference. *International conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 181–190.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S. & Li, J. (2021c). Transbts: Multimodal brain tumor segmentation using transformer. *International conference of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 109–119.

Wang, Y., Chen, H., Fan, Y., Sun, W., Tao, R., Hou, W., Wang, R., Yang, L., Zhou, Z., Guo, L.-Z. et al. (2022). Usb: A unified semi-supervised learning benchmark for classification. *Advances in Neural Information Processing Systems*, 35, 3938–3961.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600–612.

Wardlaw, J. M., Smith, E. E., Biessels, G. J., Cordonnier, C., Fazekas, F., Frayne, R. & Dichgans, M. (2013). Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration. *The Lancet Neurology*, 12(8), 822–838.

Wei, C., Mangalam, K., Huang, P.-Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A. & Feichtenhofer, C. (2023). Diffusion models as masked autoencoders. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16284–16294.

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., Chang, S. M. et al. (2010). Updated response assessment criteria for high-grade gliomas: Response Assessment in Neuro-Oncology Working Group. *Journal of Clinical Oncology*, 28(11), 1963–1972.

Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109–130.

Wolleb, J., Bieder, F., Sandkühler, R. & Cattin, P. C. (2022). Diffusion models for medical anomaly detection. *International Conference on Medical image computing and computer-assisted intervention*, pp. 35–45.

Wu, F. & Zhuang, X. (2020). CF distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 39(12), 4274–4285.

Wu, K., Du, B., Luo, M., Wen, H., Shen, Y. & Feng, J. (2019). Weakly Supervised Brain Lesion Segmentation via Attentional Representation Learning. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Proceedings*, 11766(Lecture Notes in Computer Science), 211–219.

Wu, M., Zhang, L., Yap, P.-T., Lin, W., Zhu, H. & Liu, M. (2023a). Structural MRI harmonization via disentangled latent energy-based style translation. *International Workshop on Machine Learning in Medical Imaging*, pp. 1–11.

Wu, M., Zhang, L., Yap, P.-T., Zhu, H. & Liu, M. (2025). Disentangled latent energy-based style translation: An image-level structural MRI harmonization framework. *Neural Networks*, 184, 107039.

Wu, Z., Chen, X., Xie, S., Shen, J. & Zeng, Y. (2023b). Super-resolution of brain MRI images based on denoising diffusion probabilistic model. *Biomedical Signal Processing and Control*, 85, 104901.

Wyatt, J., Leach, A., Schmon, S. M. & Willcocks, C. G. (2022a, June). AnoDDPM: Anomaly Detection With Denoising Diffusion Probabilistic Models Using Simplex Noise. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops*, pp. 650-656.

Wyatt, J., Leach, A., Schmon, S. M. & Willcocks, C. G. (2022b). ANODDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops*, pp. 650–656.

Xu, H., Othman, S. F. & Magin, R. L. (2008). Monitoring tissue engineering using magnetic resonance imaging. *Journal of bioscience and bioengineering*, 106(6), 515–527.

Yan, C., Zhang, S., Liu, Y., Pang, G. & Wang, W. (2023). Feature prediction diffusion model for video anomaly detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5527–5537.

Yan, X., Zhang, H., Xu, X., Hu, X. & Heng, P.-A. (2021). Learning semantic context from normal samples for unsupervised anomaly detection. *Proceedings of the AAAI conference on artificial intelligence*, 35(4), 3110–3118.

Yang, C., Guo, X., Chen, Z. & Yuan, Y. (2022). Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79, 102457.

Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E. & Antani, S. K. (2023). Assessing inter-annotator agreement for medical image segmentation. *IEEE Access*, 11, 21300–21312.

Yao, H., Liu, M., Wang, H., Yin, Z., Yan, Z., Hong, X. & Zuo, W. (2024a). GLAD: Towards Better Reconstruction with Global and Local Adaptive Diffusion Models for Unsupervised Anomaly Detection. *European Conference on Computer Vision*.

Yao, X., Li, R., Qian, Z., Wang, L. & Zhang, C. (2024b). Hierarchical Gaussian Mixture Normalizing Flow Modeling for Unified Anomaly Detection. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*.

Yi, J. & Yoon, S. (2020). Patch svdd: Patch-level svdd for anomaly detection and segmentation. *Proceedings of the Asian conference on computer vision*.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y. & Le, X. (2022). A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35, 4571–4584.

Yu, J., Oh, H. & Yang, J. (2023). Adversarial Denoising Diffusion Model for Unsupervised Anomaly Detection. *Deep Generative Models for Health Workshop NeurIPS 2023*.

Zang, C. & Wang, F. (2020). MoFlow: an invertible flow model for generating molecular graphs. *Proceedings of the 26th ACM SIGKDD*, pp. 617–626.

Zavrtanik, V., Kristan, M. & Skočaj, D. (2021). Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8330–8339.

Zavrtanik, V., Kristan, M. & Skočaj, D. (2022). Dsr–a dual subspace re-projection network for surface anomaly detection. *European conference on computer vision*, pp. 539–554.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., Initiative, A. D. N. et al. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3), 856–867.

Zhang, J., Chen, X., Wang, Y., Wang, C., Liu, Y., Li, X., Yang, M.-H. & Tao, D. (2025). Exploring plain ViT features for multi-class unsupervised visual anomaly detection. *Computer Vision and Image Understanding*, 104308.

Zhang, L. et al. (2020). Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE TMI*, 2531–2540.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 586–595.

Zhang, X., Li, N., Li, J., Dai, T., Jiang, Y. & Xia, S.-T. (2023a). Unsupervised surface anomaly detection with diffusion probabilistic model. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6782–6791.

Zhang, X., Li, S., Li, X., Huang, P., Shan, J. & Chen, T. (2023b). Destseg: Segmentation guided denoising student-teacher for anomaly detection. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3914–3923.

Zhang, Y., Liu, H. & Hu, Q. (2021). Transfuse: Fusing transformers and cnns for medical image segmentation. *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part I 24*, pp. 14–24.

Zhao, H., Cai, H. & Liu, M. (2024). Transformer based multi-modal MRI fusion for prediction of post-menstrual age and neonatal brain development analysis. *Medical Image Analysis*, 94, 103140.

Zhao, X. & Zhao, X.-M. (2021). Deep learning of brain magnetic resonance images: A brief review. *Methods*, 192, 131–140.

Zhao, Y., Ding, Q. & Zhang, X. (2023). AE-FLOW: Autoencoders with normalizing flows for medical images anomaly detection. *International Conference on Learning Representations (ICLR)*.

Zhao, Z., Xu, K., Li, S., Zeng, Z. & Guan, C. (2021). MT-UDA: Towards Unsupervised Cross-modality Medical Image Segmentation with Limited Source Labels. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 293–303.

Zhu, J.-Y. et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, pp. 2223–2232.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W. & Wang, X. (2024). Vision mamba: Efficient visual representation learning with bidirectional state space model. *International Conference on Machine Learning (ICML)*.

Zimmerer, D., Kohl, S. A., Petersen, J., Isensee, F. & Maier-Hein, K. H. (2018). Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*.

Zimmerer, D., Isensee, F., Petersen, J., Kohl, S. & Maier-Hein, K. (2019). Unsupervised anomaly localization using variational auto-encoders. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pp. 289–297.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301–320.

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells III, W. M., Jolesz, F. A. & Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2), 178–189.

Zou, Y., Jeong, J., Pemula, L., Zhang, D. & Dabeer, O. (2022). Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. *European Conference on Computer Vision*, pp. 392–408.

Zuo, L., Liu, Y., Xue, Y., Dewey, B. E., Remedios, S. W., Hays, S. P., Bilgel, M., Mowry, E. M., Newsome, S. D., Calabresi, P. A. et al. (2023). HACA3: A unified approach for multi-site MR image harmonization. *Computerized Medical Imaging and Graphics*, 109, 102285.

Zuo, L. et al. (2021). Information-based disentangled representation learning for unsupervised MR harmonization. *IPMI*, pp. 346–359.