UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Towards Robust Clinical Segmentation of Paediatric Brain Tumours in Magnetic Resonance Images Using Weakly-supervised Deep Neural Networks

by Nico Loesch

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Paul J. Kennedy and Prof. Daniel R. Catchpoole.

University of Technology Sydney
Faculty of Engineering & IT

January 2026

Certificate of Original Authorship

I, Nico Loesch, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

Signature: Production Note: Signature removed prior to publication.

[Nico Loesch]

Date: 28 January 2026

Place: Sydney, Australia

Acknowledgements

First of all, I would like to thank my supervisors, Prof. Paul Kennedy and Prof. Daniel Catchpoole, for their constant support and guidance throughout my PhD journey. They always reminded me to believe in myself and my accomplishments, and for that I am very grateful. I am also thankful to have been the recipient of the International Research Training Program Scholarship, which provided me with the financial support to pursue my research.

I would like to thank my friends Adrian, Tobi, and Dominic for the banter at lunch, and the laughs we shared. Your experience and insights were invaluable, especially at the start of my candidature when everything felt new and overwhelming, and it was reassuring to know I was not facing these challenges alone.

A special thanks goes to my clinical collaborators, Dr. Robert Goetti, Dr. Dinisha Govender, and Prof. Stewart Kellie, for providing valuable insights into the clinical aspects of my research. This research was also made possible through data and samples provided by the The Children's Brain Tumor Network (CBTN).

To my family, I would like to give heartfelt thanks: to my Mom, Dad, Sister, and Brother back home for their constant love and encouragement, even from afar, and to my family here in Australia for making me feel at home and supporting me throughout this journey.

And finally, my deepest gratitude goes to my partner Emma. Your unwavering support, patience, and understanding made this journey possible. Thank you for listening to my endless rants, letting me vent and complain when things got tough, and for always encouraging me to keep going. Thank you for having my back - I could not have done this without you.

Research Output

List of peer-reviewed Publications

Publication 1 **Nico Loesch**, Daniel R. Catchpoole, and Paul J. Kennedy. Three-Dimensional Latent Diffusion Model for Weakly-Supervised Brain Tumour Segmentation. In ***Artificial Intelligence in Medicine (AIME) 2025, Lecture Notes in Computer Science***, vol 15734, 2025. doi: 10.1007/978-3-031-95838-0_24. URL https://doi.org/10.1007/978-3-031-95838-0_24

Publication 2 **Nico Loesch**, Daniel R. Catchpoole, and Paul J. Kennedy. Weakly-Supervised Brain Tumour Segmentation via Latent Diffusion Models: Evaluating the Role of Super-Resolution in Anomaly Detection. ***In submission***, 2026.

Research Framework Repository

In addition to the listed publications, a comprehensive software framework was developed. The repository integrates all components of the proposed approach, including dataset preprocessing, model training and evaluation pipelines. It provides a unified and extensible codebase for reproducible experimentation with weakly-supervised latent diffusion models for brain tumour segmentation. The repository is publicly available at https://github.com/nicoloesch/ADiff.

Contents

List of Figures

List of Tables

List of Algorithms

List of Acronyms

ACL affine coupling layer
AE autoencoder
BraTS brain tumor segmentation
CAD computer-aided diagnosis
CAM class activation map
CBTN The Children's Brain Tumor Network
CE cross-entropy
CHOP The Children's Hospital of Philadelphia
CHW The Children's Hospital at Westmead
CNN convolutional neural network
CNS central nervous system
DDIM denoising diffusion implicit model
DDPM denoising diffusion probabilistic model
DICOM Digital Imaging and Communications in Medicine
DL deep learning
DMG diffuse midline glioma
DSC Dice similarity coefficient
ED peritumoral edematous and invaded tissue
EDICT exact diffusion inversion via coupled transformations
EMA exponential moving average
ESRGAN Enhanced Super-Resolution Generative Adversarial Network
ET Gadolinum-enhancing tumour
FAD Fréchet autoencoder distance
FID Fréchet inception distance
GAN generative adversarial network
HGG high-grade glioma
HR high-resolution
HREC human research ethics committee
IoU intersection over union
IS inception score
KL Kullback-Leibler

KLAE autoencoder with Kullback-Leibler regularisation
LDM latent diffusion model
LGG low-grade glioma
LPIPS learned perceptual image patch similarity
LR low-resolution
ML machine learning
MRI magnetic resonance imaging
MS-SSIM multi-scale structural similarity index
MSE mean-squared error
NCR necrotic tumour core
NIfTI Neuroimaging Informatics Technology Initiative
NLL negative log-likelihood
NN neural network
PE position embedding
PI principal investigator
PSNR peak signal-to-noise ratio
RECIST Response Evaluation Criteria in Solid Tumors
SPADE SPatially-Adaptive (DE)normalization
SR super-resolution
SSIM structural similarity index
$\boldsymbol{T}_1$**ce** T_1 contrast-enhanced
$\boldsymbol{T}_1$**w** T_1-weighted
$\boldsymbol{T}_2$**f** T_2-fluid attenuated inversion recovery
$\boldsymbol{T}_2$**w** T_2-weighted
TCIA the cancer imaging archive
UTS The University of Technology Sydney
VAE variational autoencoder
VLB variational lower bound
VQ vector quantisation
VQ-GAN vector quantized generative adversarial network
VQAE autoencoder with vector quantisation regularisation
WHO World Health Organization

Abstract

The delineation of paediatric brain tumours in magnetic resonance imaging (MRI) remains a major clinical challenge due to complex tumour presentations, and the heightened sensitivity of developing brains to treatment. Manual annotation is time-consuming and subjective, with inconsistencies arising from differences in imaging quality and tumour morphology. Automated approaches show promise in addressing these issues, yet state-of-the-art supervised deep learning (DL) methods depend on extensive, pixel-level annotations that are costly and scarce in paediatric populations. To overcome these limitations, this thesis investigates weakly-supervised anomaly detection based on denoising diffusion probabilistic models (DDPMs) as an alternative for delineating paediatric brain tumours in MRI.

The first contribution introduces a 3D-latent diffusion model (LDM) with a novel patch-based training strategy that enables efficient learning on volumetric data while reducing computational demand. This strategy also facilitates the extraction of pseudo-healthy anatomy from diseased individuals, mitigating data collection requirements. The applicability of a novel encoding mechanism, adapted from natural images, is further assessed for medical imaging. The approach surpasses existing weakly-supervised baselines across several benchmarks. However, the generation of artefacts raises important questions regarding performance on small lesion detection.

To address these limitations, the second contribution exploits the generative capacity of LDMs to synthesise datasets with precisely controlled lesion sizes. Various conditioning strategies are systematically compared to balance fidelity and dataset consistency. Building on this foundation, the third contribution explores spatial resolution enhancement via super-resolution (SR) using a conditional LDM to improve

the detection of small lesions. The results demonstrate clear gains in lesion sensitivity and resolution-aware segmentation.

The fourth contribution assesses the generalisability of the proposed framework to paediatric tumours, an underexplored domain limited by scarce annotations. Experimental results show that LDMs trained solely on adult data generalise effectively to paediatric cases, while fine-tuning yields negligible gains. Additional evaluation on a private multi-institutional cohort encompassing diverse tumour types and acquisition conditions further supports the framework's robustness. These findings demonstrate that anomaly detection can extend beyond its original training domain and underscore the framework's relevance in low-annotation regimes.

Together, these contributions advance the use of LDMs for weakly-supervised anomaly detection in medical imaging, unifying lesion detection, spatial resolution enhancement, and synthetic data generation. The framework reduces dependence on large annotated datasets and demonstrates robustness across adult and paediatric cohorts. As a result, this thesis outlines a pathway towards scalable and clinically applicable paediatric tumour segmentation beyond conventional supervised paradigms.

Chapter 1

Introduction

Paediatric central nervous system (CNS) tumours are the second most common paediatric malignancy (Hossain et al., 2021; Siegel et al., 2024). Among all CNS neoplasms, brain tumours are associated with disproportionately high mortality rates and remain a substantial health concern. Research on brain tumours has largely focused on adult populations, yet paediatric tumours exhibit distinct clinical and biological characteristics that complicate their diagnosis and management (d'Amati et al., 2024; Louis et al., 2021; Pfister et al., 2022; Siegel et al., 2024). In response, recent efforts have focused on refining stratification and treatment. The latest World Health Organization (WHO) classification (***Central Nervous System Tumours***, 2021) reflects this shift by introducing a more differentiated and biologically grounded categorisation of paediatric brain tumours.

Magnetic resonance imaging (MRI) plays a pivotal role in the diagnostic and therapeutic workflow, providing the primary modality for non-invasive tumour characterisation (d'Amati et al., 2024; Sturm et al., 2017; Villanueva-Meyer et al., 2017). As the first step in the clinical decision-making process, accurate imaging interpretation is essential: subsequent treatment planning, surgical strategy, and radiotherapy are all reliant on precise spatial delineation of the lesion to preserve healthy tissue (Carrete et al., 2022; J. Huang et al., 2022; Litjens et al., 2017; Najjar, 2023; Sterzing et al., 2011). However, the manual delineation of brain tumours, also known as segmentation, is a time-intensive task that remains vulnerable to variability between experts, diagnostic fatigue, and challenges posed by diffuse or

subtle lesions. These complexities motivate the development of assistive systems that support radiologists in the detection and analysis of tumoural regions, enhancing consistency, reducing workload, and ultimately contributing to improved clinical outcomes (Bakas et al., 2017; J. Huang et al., 2022; Willemink et al., 2020).

Deep learning (DL) has shown substantial success in brain tumour segmentation, automating the delineation of tumoural regions with high accuracy (Isensee et al., 2021; Litjens et al., 2017; Lundervold & Lundervold, 2019; R. Wang et al., 2022). However, state-of-the-art methods typically rely on costly fine-grained expert annotations. In contrast, weak labels, such as image- or volume-level annotations, are considerably easier to obtain than pixel- or voxel-wise segmentations. As a result, their use enables the integration of large, unlabelled clinical datasets into model training, facilitating broader coverage of disease variability and reducing the reliance on time-intensive manual annotation (Cheplygina et al., 2019; Kazerouni et al., 2023; Litjens et al., 2017; R. Wang et al., 2022).

The significance of the research in this thesis is threefold. Firstly, it builds on the established body of work in weakly-supervised brain tumour segmentation and advances it by incorporating volumetric MRI data. This enables richer spatial context and improved lesion localisation, while necessitating architectural adaptations for efficient feature extraction in high-dimensional settings.

Secondly, this research addresses the challenge of early detection through the surrogate objective of small lesion segmentation. Early-stage tumours are inherently small and often subtle, yet timely identification is essential in paediatric neuro-oncology to prevent long-term neurological and developmental impairments. By systematically evaluating performance depending on lesion size using synthetic datasets, this work provides a controlled setting to quantify detectability under clinically relevant conditions. In addition, it investigates the impact of increased resolution on the detectability of small lesions, providing insights into the trade-offs between image quality, computational cost, and segmentation performance.

Finally, this research builds on both preceding contributions to explore a pathway towards the segmentation of paediatric brain tumours, leveraging insights gained in

the adult domain. It examines the transfer of representations learned on adult imaging data to the paediatric setting, where research remains scarce due to the absence of large, annotated datasets. In this context, it further evaluates the generalisability of the proposed approach, aiming to reduce reliance on limited annotations and improve resilience to distributional shifts. This strategy lays the groundwork for scalable weakly-supervised frameworks applicable across both adult and paediatric populations.

1.1 Paediatric brain tumours

Paediatric CNS tumours are the most common solid tumours in children, accounting for 25% of all childhood cancers and 21% of cancers in adolescents (Siegel et al., 2024). These tumours are often aggressive, particularly in younger children, where only one-third are benign, contributing to a high mortality rate of 35-40% (Hossain et al., 2021; Siegel et al., 2024). Due to the ever-changing landscape of these lesions, the WHO periodically releases classification and categorisation guidelines. Generally speaking, the lesions are classified based on their ***location*** within the brain, their ***appearance in medical imaging*** and their ***molecular profile and histological features*** obtained through biopsy. Particularly molecular diagnostics have become increasingly important for accurate classification and are reflected in the most recent 5th edition of the "WHO Classification of Central Nervous System" (***Central Nervous System Tumours***, 2021; d'Amati et al., 2024; Louis et al., 2021; Pfister et al., 2022). A major change in these guidelines is the introduction of distinct subcategories for paediatric tumours to reflect their fundamental divergence from adult counterparts. Unlike adult cancers, which typically result from long-term environmental exposure and the gradual accumulation of genetic damage, paediatric tumours are often the result of "maturation blocks" during early development (Behjati et al., 2021; Pui et al., 2011). Rather than having a high number of genetic mutations gathered over decades, childhood tumours usually feature far fewer genetic changes, often driven by a single major event (e.g. a specific gene fusion) that occurs in immature cells

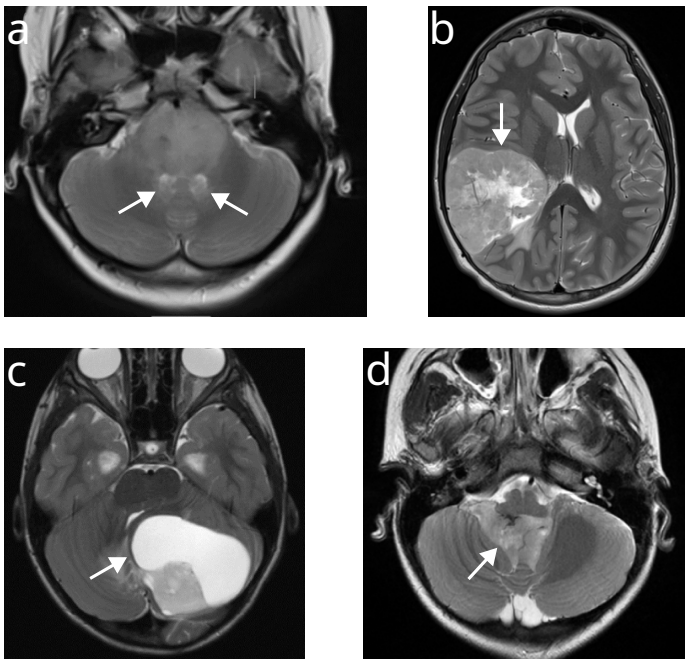


Figure 1.1: Axial T_2-weighted MRIs of paediatric brain tumours: (a) diffuse midline glioma, (b) high-grade glioma, (c) low-grade glioma, and (d) ependymoma. Arrows indicate the tumour. *Note.* All images are adapted from Jaju et al. (2022). CC BY 4.0.

(Gröbner et al., 2018; Ma et al., 2018). Additionally, these tumours are less likely to be detected or attacked by the body's natural immune system compared to their adult counterparts (Grabovska et al., 2020; Wienke et al., 2021).

Paediatric brain tumours are primarily classified into gliomas and ependymal tumours. Gliomas, the most common subtype accounting for approximately 50% of new diagnoses, originate from glial cells and are categorised into ***low-grade gliomas (LGGs)*** and ***high-grade gliomas (HGGs)*** based on malignancy and molecular features (d'Amati et al., 2024; Pfister et al., 2022; Wells & Packer, 2015; Zong et al., 2015). LGGs often occur in the cerebellum, optic pathway, and brainstem and exhibit slow growth associated with favourable prognosis (Banerjee & Nicolaides, 2017; d'Amati et al., 2024; Pfister et al., 2022; Wells & Packer, 2015). In contrast, HGGs, including diffuse midline gliomas (DMGs), demonstrate rapid progression and poor survival rates, partially due to their location in indispensable brain compartments including cerebellum, thalamus, and pons (J. S. Chang et al., 2017; d'Amati et al.,

2024; Pfister et al., 2022; Wells & Packer, 2015). Ependymal tumours, the third most common subtype, arise from CNS ventricular lining cells and are classified based on location into supratentorial and posterior fossa ependymomas. Pediatric ependymomas, classified into WHO grades I - III, exhibit heterogeneous prognosis, with grade I lesions generally associated with favourable outcomes, while grade III (anaplastic) ependymomas carry a markedly worse prognosis (d'Amati et al., 2024; Kline et al., 2017; Wells & Packer, 2015).

1.1.1 Conventional diagnosis and treatment

Depending on lesion location, tumour biology, and patient age, paediatric brain tumours can present with a wide range of clinical symptoms. These are typically classified as either non-specific or focal. Non-specific symptoms include headache, nausea, vomiting, behavioural changes, and subtle developmental delays. Focal symptoms are more directly linked to the lesion's anatomical location and may manifest as speech impairments, visual disturbances, or motor deficits such as hemiparesis (Fry et al., 2014; Lanphear & Sarnaik, 2014; Wells & Packer, 2015; Wilne et al., 2010; Yamada et al., 2020; Zumel-Marne et al., 2020).

Symptom severity often correlates with tumour grade: higher grade lesions tend to provoke rapid and severe symptoms, prompting earlier clinical investigation. In contrast, low-grade tumours are more likely to cause mild and non-specific symptoms, frequently resulting in delayed diagnosis (Banerjee & Nicolaides, 2017; J. S. Chang et al., 2017; d'Amati et al., 2024).

The importance of magnetic resonance imaging

The radiologic intervention is the cornerstone of the diagnostic pathway for patients with suspected brain tumours, representing the critical first step in their clinical journey. As the primary imaging modality, MRI is preferred for its superior soft-tissue contrast, absence of ionising radiation, and multi-sequence capability, enabling detailed characterisation of various disease characteristics and guiding subsequent diagnostic and therapeutic decisions (Carrete et al., 2022; Thust et al., 2018;

Villanueva-Meyer et al., 2017). Beyond tumour detection, radiological assessment provides crucial information on lesion location, morphological features, and tissue composition, which facilitates a preliminary classification that informs treatment strategies. This imaging data is fundamental to precise, patient-specific treatment while minimising toxicity and preserving healthy tissue (Mueller & Chang, 2009; Vagvala et al., 2022; Villanueva-Meyer et al., 2017; Wells & Packer, 2015).

The recommended MRI clinical protocol includes T_1-weighted (T_1w) and T_1 contrast-enhanced (T_1ce) sequences, along with T_2-weighted (T_2w) and T_2-fluid attenuated inversion recovery (T_2-FLAIR) sequences. Increasingly, 3D sequences are replacing traditional 2D planar sequences due to their improved anatomical fidelity and enhanced through-plane resolution, despite the longer acquisition times (Bakas et al., 2017; Carrete et al., 2022; Thust et al., 2018; Villanueva-Meyer et al., 2017). Standard clinical MRI protocols typically operate at a spatial resolution (voxel size) in the order of $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ for isotropic 3D acquisitions, or approximately $1\,\text{mm} \times 1\,\text{mm} \times 5\,\text{mm}$ for traditional 2D axial slices (Bakas et al., 2017; Thust et al., 2018). Understanding these dimensions is critical when considering the targeted lesion sizes in this research and becomes particularly relevant when evaluating the detectability of small lesions (see Section 2.1.6 and Chapter 4).

T_1W MRI sequences are particularly useful to highlight blood products and calcifications, whereas T_2w sequences are more sensitive to the presence of oedema, cystic components, and necrosis. These features allow the initial structural classification of the lesion. LGGs are generally more circumscribed and often present with both cystic and solid components. In contrast, HGGs typically exhibit diffuse infiltration of surrounding structures, with associated oedema, increased vascularity, and disruption of the blood-brain barrier (Banerjee & Nicolaides, 2017; Carrete et al., 2022; J. S. Chang et al., 2017; Kline et al., 2017; Thust et al., 2018; Villanueva-Meyer et al., 2017). T_2-FLAIR sequences help evaluate the extent of lesions, particularly in cases with oedema or lower-grade lesions that exhibit low contrast enhancement on T_1w images. This capability aids in assessing affected brain regions and delineating lesion borders (Carrete et al., 2022; Thust et al., 2018; Villanueva-Meyer et al., 2017).

In addition, advanced techniques such as perfusion- and diffusion-weighted imaging are increasingly being incorporated. These sequences provide valuable insights into tumour vascularity and cellularity, respectively, offering critical information for lesion characterisation and treatment planning. However, due to their higher cost and lack of standardisation in clinical protocols, these advanced imaging modalities are predominantly employed in research settings rather than routine diagnostics (Carrete et al., 2022; Thust et al., 2018; Villanueva-Meyer et al., 2017).

Standard treatment regimens

Conventional MRI assessment is followed by a biopsy to confirm the diagnosis and provide a histological and molecular classification of the tumour. The combination of imaging data and tumour profiling is critical as it determines the subsequent treatment strategy. With the ongoing development of novel therapies targeting newly identified molecular markers, the treatment landscape is becoming increasingly complex. In general, the primary aim of the treatment is to balance effectiveness and potential curative approaches with the preservation of normal neurological function (Carrete et al., 2022; Eisenhauer et al., 2009; J. Huang et al., 2022; Mueller & Chang, 2009; Sterzing et al., 2011; Vagvala et al., 2022; Wells & Packer, 2015).

Treatment is typically initiated with surgical resection, aiming for gross-total resection, which can eliminate the need for adjuvant therapy and is associated with excellent long-term survival outcomes (Banerjee & Nicolaides, 2017; J. S. Chang et al., 2017; d'Amati et al., 2024; Kline et al., 2017). However, complete resection is often impractical for deep-seated or diffusely infiltrative lesions due to the high risk of permanent neurological deficits. In such cases, alternative or adjuvant treatment strategies are employed, including radiation therapy, chemotherapy, and targeted therapies. Radiation therapy is commonly used in cases of recurrence or for older patients but is generally avoided in younger populations due to the risk of neurocognitive impairment and secondary malignancies (J. S. Chang et al., 2017; Kline et al., 2017; Wells & Packer, 2015). Advances such as proton beam therapy and brachytherapy have been developed to minimise off-target radiation exposure

and reduce long-term toxicity. Chemotherapy may complement or replace surgery and radiation therapy, particularly for systemic tumour control or in instances where surgical intervention is not viable; however, toxicity considerations remain a major limitation (Banerjee & Nicolaides, 2017; J. S. Chang et al., 2017; d'Amati et al., 2024; Vagvala et al., 2022; Wells & Packer, 2015). More recently, targeted therapies, including molecular inhibitors and immunotherapy, have been explored to improve treatment outcomes and reduce therapy-associated toxicity. These approaches seek to disrupt tumour-specific pathways involved in proliferation and invasion or enhance the immune system's ability to recognise and eliminate malignant cells (Banerjee & Nicolaides, 2017; J. S. Chang et al., 2017; d'Amati et al., 2024; Kline et al., 2017; Vagvala et al., 2022; Wells & Packer, 2015).

Challenges of conventional diagnosis and treatment

Human analysis remains the gold standard in the diagnostic evaluation of paediatric brain tumours, with radiologists playing a central role in identifying and interpreting abnormalities (Thust et al., 2018; Villanueva-Meyer et al., 2017). However, the diagnostic pathway is inherently complex, influenced by both technical limitations of imaging modalities and the challenges associated with human interpretation. MRI, while offering superior soft-tissue contrast and multi-sequence capability, is subject to resolution constraints, imaging artefacts, and signal-to-noise limitations, all of which may hinder the detection of subtle or small lesions. These technical constraints, coupled with the variable presentation of paediatric tumours, contribute to the difficulty of early and accurate diagnosis (Banerjee & Nicolaides, 2017; Bruno et al., 2015; Carrete et al., 2022; J. S. Chang et al., 2017; Goldman et al., 2017; Villanueva-Meyer et al., 2017).

Beyond these modality-specific challenges, human interpretation is influenced by perceptual and cognitive factors. Perceptual errors occur when abnormalities are overlooked during image evaluation, often exacerbated by clinician fatigue, high workloads, or the difficulty in identifying secondary findings once an initial diagnosis is made. Cognitive errors, on the other hand, involve incorrect interpretation of

detected abnormalities, including misclassification of lesion significance or diagnostic anchoring based on incomplete clinical information (Bruno et al., 2015; Lee et al., 2013; L. Zhang, Wen, et al., 2023). The interplay of these factors contributes to the potential for misdiagnosis, which can substantially affect patient outcomes by delaying treatment initiation, leading to increased neurological morbidity, long-term functional impairment, and, in some cases, poorer survival rates. This impact is particularly pronounced in paediatric cases, where delays in intervention may irreversibly affect critical developmental windows (Banerjee & Nicolaides, 2017; Bruno et al., 2015; J. S. Chang et al., 2017; Goldman et al., 2017; Sadighi et al., 2018; Smith et al., 2008; Sturm et al., 2017; Wells & Packer, 2015).

1.2 How can machine learning help?

Despite the tremendous efforts employed in clinical practice to address the challenges inherent in diagnosing paediatric brain tumours, misdiagnosis rates continue to reach up to 30% (Bruno et al., 2015; Goldman et al., 2017; Lee et al., 2013; L. Zhang, Wen, et al., 2023). Accurate tumour measurement is critical not only for early detection but also for assessing disease progression and evaluating the efficacy of treatment over time. Given the complexity of tumour identification, particularly with small or subtle lesions, and the potential for errors in human interpretation, there is a growing need for alternative solutions to assist medical professionals in their assessment processes (Bakas et al., 2017; Najjar, 2023; R. Wang et al., 2022; Zegers et al., 2021).

Computer-aided diagnosis (CAD) systems offer promising avenues for improving diagnostic accuracy. These systems aim to support radiologists by providing objective and consistent analyses of medical images through advanced image processing techniques (Najjar, 2023; R. Wang et al., 2022; Zegers et al., 2021). CAD systems can help mitigate perceptual and cognitive errors by directing clinician attention to regions that warrant further investigation. Additionally, these systems can assist in automating tasks such as tumour segmentation, which forms the basis for volumetric measurement and longitudinal tracking. Both are crucial for treatment planning as

well as for assessing treatment efficacy and disease progression. As such, integrating CAD systems into clinical workflows could provide a complementary tool for clinicians, enhancing diagnostic precision and reducing the potential for human error in the interpretation of complex imaging data (Bakas et al., 2017; Litjens et al., 2017; R. Wang et al., 2022).

The success of DL in pattern recognition tasks across various domains has led to an increased focus on transferring these methodologies to the medical field. Particularly the ability to automatically detect patterns in complex data using convolutional neural networks (CNNs) is beneficial in the medical domain, where the identification of disease-related patterns in digital medical imaging data is crucial (Isensee et al., 2021; Krizhevsky et al., 2012; Litjens et al., 2017; Ronneberger et al., 2015). As such DL-based CAD systems are actively researched and developed to assist clinicians in the diagnosis of various diseases, including brain tumours. These systems aim to enhance diagnostic accuracy, reduce misdiagnosis rates, and improve overall patient outcomes by offering clinicians more precise and detailed insights into patient conditions (Isensee et al., 2021; Litjens et al., 2017; Lundervold & Lundervold, 2019; Painuli et al., 2022; R. Wang et al., 2022; H. Yu et al., 2021).

For the diagnosis of brain tumours from digital MRI scans, DL primarily addresses two tasks: classification and segmentation. Classification assigns scans to tumour types by extracting discriminative features, supporting diagnosis through identification of class-specific pathological patterns (Litjens et al., 2017; H. Yu et al., 2021). Segmentation delineates tumour boundaries at the pixel level, providing spatial information on size, shape, and location (Eisenhauer et al., 2009; Litjens et al., 2017; Najjar, 2023; R. Wang et al., 2022; H. Yu et al., 2021). By embedding tumour characteristics within their anatomical context, segmentation complements classification and enables precise clinical assessment (Bishop & Bishop, 2024a; Lundervold & Lundervold, 2019; Painuli et al., 2022; R. Wang et al., 2022; H. Yu et al., 2021). This thesis therefore focuses on segmentation, as accurate delineation underpins treatment planning, surgical decision-making, and longitudinal monitoring.

1.2.1 Supervised brain tumour segmentation

Supervised DL has been instrumental in the success of automatic pattern detection across various domains. It represents the gold standard in medical imaging analysis due to its efficiency and accuracy (Najjar, 2023; R. Wang et al., 2022; H. Yu et al., 2021). In supervised learning, the training process is guided by expert-provided annotations that serve as reference targets. These supervisory signals constrain the optimisation process by narrowing the search space and enabling the model to learn task-relevant patterns within complex data. In medical imaging, these techniques have proven particularly effective for tasks such as segmentation and lesion detection, offering considerable potential to assist clinicians in tumour diagnosis and treatment planning (Najjar, 2023; R. Wang et al., 2022; H. Yu et al., 2021).

Limitations of supervised approaches

While supervised DL has demonstrated considerable success in domains with abundant annotated data, this strength becomes a major limitation in medical imaging, where data for specific conditions and populations is often scarce. Although one might expect substantial volumes of medical imaging data for brain tumour patients, given the high number of new diagnoses and radiological assessments conducted annually (Miller et al., 2021; Siegel et al., 2019, 2024), access to such datasets is severely restricted due to ethical and privacy concerns (Kush et al., 2020; Varoquaux & Cheplygina, 2022; Willemink et al., 2020; Wirth et al., 2021). This issue is particularly pronounced in paediatric imaging, where data sharing is further constrained due to the sensitive nature of the population, thereby limiting the availability of data for specific paediatric aetiologies (Patrinos et al., 2022; Price & Cohen, 2019).

Even if ethical barriers to data access are overcome, supervised DL still requires high-quality datasets, encompassing both imaging data and expert-level annotations. The manual labelling of medical images to obtain the ground truth required for supervised training is a labour-intensive and costly process that demands expert input at every stage. This is particularly important for segmentation tasks, which

require voxel-wise annotations of multimodal and 3D imaging data to ensure accurate lesion delineation. The availability of such annotated datasets is severely constrained with data often limited to specific pathologies, imaging modalities, and patient demographics (Adewole et al., 2023; Kazerooni et al., 2024; Moawad et al., 2024). For example, until 2023, the leading dataset in the brain tumor segmentation (BraTS) challenge was limited to adult gliomas, with only recently added paediatric cases and expanded pathology coverage (Adewole et al., 2023; Kazerooni et al., 2024; Moawad et al., 2024). In addition, the annotation process is further complicated by the quality of the raw imaging data, which is subject resolution constraints, imaging artefacts, and signal-to-noise issues (Aja-Fernández & Vegas-Sánchez-Ferrero, 2016; Bruno et al., 2015; Goldman et al., 2017; Soomro et al., 2023; Willemink et al., 2020).

Furthermore, the reliance on expert annotations introduces another critical challenge: the potential for bias and variability in the training process. Annotation quality can vary between experts due to differences in experience, interpretation, and institutional protocols, leading to inconsistencies within the training data. These discrepancies may propagate through the learning process, adversely impacting model performance and reducing its ability to generalise effectively, particularly when applied to external datasets or previously unseen pathologies (Bruno et al., 2015; Goldman et al., 2017; Lee et al., 2013; Varoquaux & Cheplygina, 2022; Willemink et al., 2020; L. Zhang, Wen, et al., 2023). As a result, even when large datasets are available, annotation inconsistencies pose an additional barrier to the development of robust supervised DL models for brain tumour segmentation.

Despite efforts to provide a gold-standard dataset, the available collections remain relatively small compared to those in natural image analysis, which often contain hundreds of thousands of samples (Deng et al., 2009; Lin et al., 2014; F. Yu et al., 2015). The largest dataset in the BraTS challenge includes only 1251 adult glioma subjects, which may not capture the full diversity of gliomas, let alone of differing brain tumour variants. Consequently, supervised DL models trained on such datasets may exhibit limited generalisability to data outside the scope of these samples, especially when focused on specific pathologies (Ali et al., 2024; J. Peng & Wang,

2021; Varoquaux & Cheplygina, 2022). This scarcity is further amplified in the paediatric domain as outlined before. The current state-of-the-art dataset contains only 99 subjects (Kazerooni et al., 2024), which is an order of magnitude smaller than the adult counterpart, and is likely of insufficient scale for training high-capacity DL models (Sun et al., 2017; Varoquaux & Cheplygina, 2022; Willemink et al., 2020).

1.2.2 Brain tumour segmentation with reduced supervision

One active research direction seeks to reduce reliance on ground-truth annotations for brain tumour segmentation by developing approaches that circumvent the limitations of fully supervised learning. A prominent strategy is ***anomaly detection***, which targets regions that deviate from healthy anatomy. Early work in this area employed reconstruction-based autoencoders (AEs), trained on scans from healthy individuals to learn a compact representation of normal anatomy. Lesions are then identified as regions with elevated reconstruction error, reflecting a mismatch between the input and the learned distribution of healthy data (Baur et al., 2021; C. Zhou & Paffenroth, 2017). Subsequent iterations of this concept utilised the capacity of generative models to estimate the underlying probability distribution of healthy data. These models allow to detect examples with differing characteristics with remarkable results and performance close to conventional supervised DL in specific circumstances (Baur et al., 2019; Pinaya, Tudosiu, et al., 2022; Schlegl et al., 2019; Weninger et al., 2019; X. Wu et al., 2021; Zimmerer et al., 2019).

Since their introduction by Ho et al. (2020), denoising diffusion probabilistic models (DDPMs) have established themselves as a distinct class of generative models, characterised by high-fidelity and diverse image generation (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021). These models benefit from an easier training process due to the absence of an adversarial training scheme, avoiding mode collapse observed in generative adversarial networks (GANs). Furthermore, their denoising-based loss function enhances reconstruction quality, mitigating the blurriness often seen in AEs (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021). As a result, DDPMs have demonstrated remarkable results in the

detection of anomalies in medical images (Fontanella et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022).

The effectiveness of DDPMs in detecting anomalies in medical imaging can be attributed to several factors. Firstly, ***weakly-supervised learning*** anomaly detection approaches have substantially enhanced performance by incorporating both healthy and anomalous data during training, outperforming DDPMs trained solely on healthy data (Fontanella et al., 2024; Sanchez et al., 2022; Wolleb et al., 2022). Secondly, encoding anatomical information within the latent space of the diffusion process, combined with classifier-guided (Dhariwal & Nichol, 2021) and classifier-free (Ho & Salimans, 2021) sampling, enables the synthetic generation of healthy counterfactuals of diseased imaging data (Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022). Lastly, the introduction of latent diffusion models (LDMs) (Rombach et al., 2022) has improved DDPM performance by compressing the input data into a lower-dimensional latent space using a pre-trained AE, reducing computational complexity (Pinaya, Graham, et al., 2022).

Additionally, DDPMs offer a flexible framework in which the generative process can be effectively guided through auxiliary conditioning signals, enabling a broad range of applications such as super-resolution (SR) reconstruction and denoising (H. Chung et al., 2023; Mao et al., 2023; J. Wang et al., 2024), as well as the synthesis of specific lesion characteristics (Dorjsembe et al., 2024; Konz et al., 2024; H. Wu et al., 2024), which have the potential to further enhance anomaly detection performance.

1.3 Problem statement

In summary, paediatric brain tumours represent a severe health challenge, where early and precise diagnosis is vital for improving outcomes, as treatment delays can substantially worsen prognosis. Accurate delineation of lesion borders is central to treatment planning, enabling the preservation of healthy tissue and reducing complications during surgery or radiation therapy. This is of particular importance in paediatric patients, where the developing brain is especially vulnerable to treatment.

Despite the importance of accurate diagnosis, the manual analysis of radiological data remains the primary clinical approach, guiding subsequent therapeutic decisions. However, this process is constrained by both technical limitations (resolution constraints, imaging artefacts, and signal-to-noise issues) and human-specific factors, such as inter-observer variability, cognitive bias, and misdiagnosis rates. **These challenges underscore the urgent need for automated tools that can assist radiologists in delineating lesion boundaries in uncertain cases.**

DL-based approaches have shown great promise in medical imaging, particularly in identifying pathological features and segmenting tumours. However, supervised learning is highly dependent on large, high-quality annotated datasets, which are scarce in medical imaging due to the labour-intensive and costly nature of expert labelling. This limitation is particularly restrictive in paediatric populations, where data availability is further constrained by ethical and privacy concerns, reducing the capacity of supervised DL.

To mitigate this issue, weakly-supervised anomaly detection has emerged as a promising alternative. In this context, DDPMs are especially relevant as they learn the underlying distribution of healthy anatomy and generate counterfactual representations. This facilitates superior anomaly detection without requiring pixel-wise annotations. In addition, their ability to operate without pixel-level labels makes them particularly well-suited to leverage large-scale unlabelled hospital data and unlock new avenues for detecting previously unseen anomalies across diverse populations. However, current weakly-supervised DDPMs lack the robustness and accuracy required to replace supervised approaches. Key challenges include the detection of small lesions, adaptation to heterogeneous imaging distributions, and ensuring reliable anomaly identification beyond predefined pathologies and populations. Addressing these limitations is essential for the clinical translation of generative models, potentially enabling early tumour detection and improved treatment planning in paediatric brain tumour patients.

1.3.1 Research Gaps and Questions

Building on the challenges outlined in Chapter 2, this research identifies three key gaps that limit the effectiveness of weakly-supervised brain tumour segmentation with DDPMs. These gaps form the basis for the research questions addressed in this thesis:

Gap 1 The transition from 2D to 3D medical image analysis for weakly-supervised brain tumour segmentation remains underexplored due to increased computational demands and the lack of public datasets with healthy individuals.

Gap 2 The role of SR in enhancing weakly-supervised DDPM-based anomaly detection remains underexplored, particularly with respect to its effect on sensitivity to small or subtle brain tumours.

Gap 3 The generalisability of weakly-supervised DDPMs to paediatric brain tumours remains untested, despite their hypothesised robustness to distributional shifts and suitability for data-scarce clinical settings.

Gap 1 and Gap 2 represent technical precursors, addressing the limited support for 3D modelling and the reduced sensitivity to small lesions, respectively. Resolving these issues is aimed to improve robustness and anomaly detection accuracy. Such improvements are essential prior to meaningful evaluation on data-scarce paediatric cohorts, as training high-capacity DDPMs requires larger datasets. In contrast, Gap 3 is application-oriented, focusing on generalisability to underrepresented paediatric populations and the challenges posed by scarce data, age-dependent anatomical variation, and clinically subtle lesion presentations. Collectively, addressing these gaps strengthens the modelling framework and establishes a pathway towards robust and scalable paediatric brain tumour segmentation.

Following the identification of these gaps, I devised the following research questions that this research project aims to systematically address:

Research Question 1 How can weakly-supervised 3D anomaly detection be made computationally efficient while mitigating reliance on publicly available datasets of healthy individuals?

The first research question aims to address Gap 1 by investigating methods for improving the efficiency and applicability of DDPMs in weakly-supervised brain tumour segmentation. It explores sampling strategies for scalable 3D processing and examines how healthy anatomical information can be obtained from diseased individuals. Additionally, it evaluates the impact of different encoding mechanisms on the quality and reliability of the learned representations, with the goal of enabling more effective anomaly detection in higher-dimensional medical imaging.

Research Question 2 What is the effect of DDPM-based SR on the sensitivity and segmentation performance of weakly-supervised anomaly detection, particularly for small brain tumours?

To address the second research question, this work investigates whether DDPM-based super-resolution improves the sensitivity of weakly-supervised anomaly detection models to small brain tumours. Following insights from Research Question 1, the investigation focuses on synthetically generating lesions with controlled properties to enable systematic evaluation. The research question further explores how SR influences segmentation performance in this high-sensitivity setting and analyses the impact of spatial resolution on the model's ability to robustly detect lesions.

Research Question 3 To what extent can DDPMs trained on adult brain tumour data generalise to the paediatric domain, and how robust are the learned representations to shifts in population and disease distribution?

A critical aspect of model robustness is its capacity to generalise across varying populations and clinical scenarios. The final research question therefore investigates how the findings from the previous research questions can be transferred to the paediatric domain, where data scarcity, anatomical variability, and clinical sensitivity

present unique challenges. By evaluating the framework on paediatric brain tumours, this thesis examines whether the proposed strategies remain effective on a distinct and more demanding population. It also proposes a staged pathway for translation towards paediatric application and seeks to demonstrate the broader applicability and adaptability of the weakly-supervised anomaly detection with DDPMs.

1.4 Contributions to knowledge

This research project makes substantial contributions to the field of weakly-supervised brain tumour segmentation using DDPMs. The contributions are structured around the three research questions outlined in Section 1.3.1.

Contribution 1 Developed a patch-based LDM for efficient 3D weakly-supervised brain tumour segmentation via anomaly detection, preserving volumetric context and integrating a robust encoding strategy to improve anatomical fidelity.

To address the computational constraints associated with volumetric data processing in 3D weakly-supervised medical image segmentation, this work introduces a DDPM-based segmentation framework that operates on spatially localised subvolumes. The framework is designed to generalise across related tasks and serves as a foundation for subsequent extensions throughout this research. A patch-based LDM is developed to enable memory-efficient training and inference while maintaining access to rich volumetric context. This design choice directly mitigates the scalability challenges outlined in Gap 1 and supports the feasibility of 3D weakly-supervised segmentation with limited computational resources. In parallel, a novel sampling mechanism is proposed to extract healthy regions from pathological volumes, circumventing the need for external control groups and enabling direct comparison to state-of-the-art approaches. Lastly, the impact of encoding strategies on the latent representation is explored through the integration of a more robust mechanism. Specifically, a strategy defined for natural imaging is explored in the context of medical imaging data, which promotes substantial conditional alterations with limited

encoding. Collectively, these components form the basis for answering Research Question 1, demonstrating that efficient 3D weakly-supervised brain tumour segmentation is achievable without compromising anatomical accuracy. This contribution is detailed in Chapter 3, with key findings published in Loesch et al. (2025).

Contribution 2 Developed and validated a DDPM-based model for the controlled synthesis of size-specific brain tumour lesions in multi-sequence MRI.

Contribution 3 Conducted the first systematic investigation of DDPM-based SR in weakly-supervised anomaly detection, evaluating its impact on the detectability of small brain tumours.

The second research project builds on the findings of the first by shifting focus to a key limitation in weakly-supervised segmentation: the detection of small lesions. These lesions are of particular clinical relevance due to their association with early-stage tumours, yet may pose major challenges for generative models due to their subtle image alterations and reduced spatial footprint. Addressing this, Contribution 2 introduces a synthetic lesion generation pipeline based on the LDM architecture established during Contribution 1. By exploring different conditioning strategies, this contribution enables the controlled synthesis of lesions with specific size and location attributes, providing a reproducible testbed for evaluating model sensitivity to small, localised anomalies. This synthetic dataset serves as a surrogate for real-world clinical cases where annotated small lesions are unavailable, thereby enabling the systematic investigation of segmentation thresholds and failure modes.

Building on this foundation, Contribution 3 explores whether increasing the spatial fidelity of input data can improve the sensitivity of weakly-supervised DDPMs to small lesions. A DDPM-based SR module is implemented using the same conditioning infrastructure as the synthetic generation task. This shared core architecture allows a seamless integration between synthetic generation and resolution enhancement by simply changing the conditioning signal. The SR module is evaluated on its ability to improve lesion visibility and segmentation accuracy in real and synthetic lesion

cases, providing a comprehensive view of its clinical applicability. By quantifying the impact of SR on boundary delineation and segmentation performance, this contribution offers empirical insight into the utility of spatial resolution enhancement in weakly-supervised frameworks, further addressing Gap 2 and answering Research Question 2. Both contributions are detailed in Chapter 4.

Contribution 4 Demonstrated the generalisability of DDPM-based weakly-supervised anomaly detection for brain tumour segmentation to data-scarce paediatric cases, validated on a curated multi-institutional dataset under clinically realistic conditions.

The last contribution extends the previously established DDPM-based framework to evaluate its applicability to paediatric brain tumour segmentation. The paediatric domain is a data-scarce setting that differs markedly from the adult population on which the earlier components of this project were developed. Building on the pretrained adult model, this contribution investigates whether its components can be leveraged to facilitate transfer of knowledge across populations. By assessing segmentation performance under distribution shift and quantifying the extent to which adaptation is required, it explores the role of conditioning in re-utilising transferable priors. This contribution addresses Gap 3 by demonstrating the framework's potential to generalise across clinically heterogeneous cohorts. By validating methods in adult cohorts and adapting them for scarce paediatric data, it provides a staged pathway towards paediatric brain tumour segmentation. This work therefore moves the framework closer to practical paediatric applications and supports broader clinical translation of the weakly-supervised DDPM approach. The findings are detailed in Chapter 5.

1.5 Significance

Brain tumours remain a particularly challenging domain in medical imaging, where timely and accurate diagnosis is critical to improving patient outcomes. This is

especially important in paediatric populations, where diagnostic delays can lead to irreversible neurological, cognitive, and developmental impairments. While DL has demonstrated impressive performance in automating brain tumour segmentation, the prevailing reliance on supervised methods imposes major limitations. These models require extensive voxel-level annotations that are time-consuming, expensive and scarce - especially in data-sensitive contexts such as paediatric brain tumours. As a result, their applicability to pediatric populations is severely restricted due to their limited capacity to generalise across clinical cohorts, imaging distributions, and lesion types.

This research addresses these limitations by advancing the field of weakly-supervised brain tumour segmentation using DDPMs. By reducing the dependence on dense annotations, this work unlocks the potential to learn from larger, more heterogeneous datasets, which aims to improve model robustness and generalisability.

Specifically, this work demonstrates how weakly-supervised anomaly detection can be extended to 3D brain imaging. This approach allows to better capture the spatial structure of tumours and improves the accuracy of lesion delineation. The significance of this work further lies in its focus on small lesion detection. Detecting subtle brain tumours is clinically vital for early-stage diagnosis but often poorly supported and analysed by existing methods. By exploring the impact of spatial resolution on the detectability of small lesions, this research provides empirical evidence for the effectiveness of increased resolution in enhancing segmentation performance in the weakly-supervised setting. Finally, the project evaluates the capacity of pre-trained models to generalise across domains by transferring them from adult to paediatric brain tumour cases. This directly supports the overarching motivation of weakly-supervised learning: to build scalable, adaptable systems that retain performance in real-world, data-scarce environments.

Collectively, this thesis advances the field of paediatric brain tumour segmentation by strengthening weakly-supervised anomaly detection as a scalable alternative to data-intensive supervised approaches. Through targeted improvements in accuracy, sensitivity, and resolution-aware detection, the work enhances the reliability of existing

diffusion-based frameworks. Crucially, it also evaluates the method's applicability to data-scarce paediatric cohorts, demonstrating promising generalisation from adult to paediatric populations. These developments mark a step toward more robust and accessible pediatric tumour segmentation pipelines. The ultimate goal is to enable earlier and more precise diagnoses, which may contribute to improved treatment planning and better outcomes for affected patients.

1.6 Thesis structure

Figure 1.2 provides an overview of the thesis structure and its alignment with the research gaps and questions identified in **Chapter 2**. This chapter reviews the current state of brain tumour segmentation, highlights limitations of supervised learning, and outlines the emerging role of weakly-supervised approaches using DDPMs. These insights form the basis for three subsequent research chapters, each addressing one of the identified gaps.

Chapter 3 introduces the core modular framework and addresses the challenges of 3D modelling for weakly-supervised segmentation. After validating the model on natural images in Section 3.2, the chapter transitions to a medical setting by proposing a patch-based 3D LDM, enabling volumetric segmentation under constrained computational resources (see Section 3.3). The chapter concludes with an evaluation of the model's performance in Section 3.4, a discussion of the limitations in Section 3.5 and a summary of key findings and contributions in Section 3.6.

Chapter 4 builds on the preceding chapter by focusing on the clinically important task of small lesion detection. Section 4.3 presents a generative pipeline for creating a synthetic small lesion dataset, which is used to systematically assess lesion detection. Section 4.4 introduces a SR module that aims to enhance spatial fidelity and examines how conditioning mechanisms and degradation strategies influence segmentation accuracy. The limitations of the approach are discussed in Section 4.5, and key findings and contributions are summarised in Section 4.6.

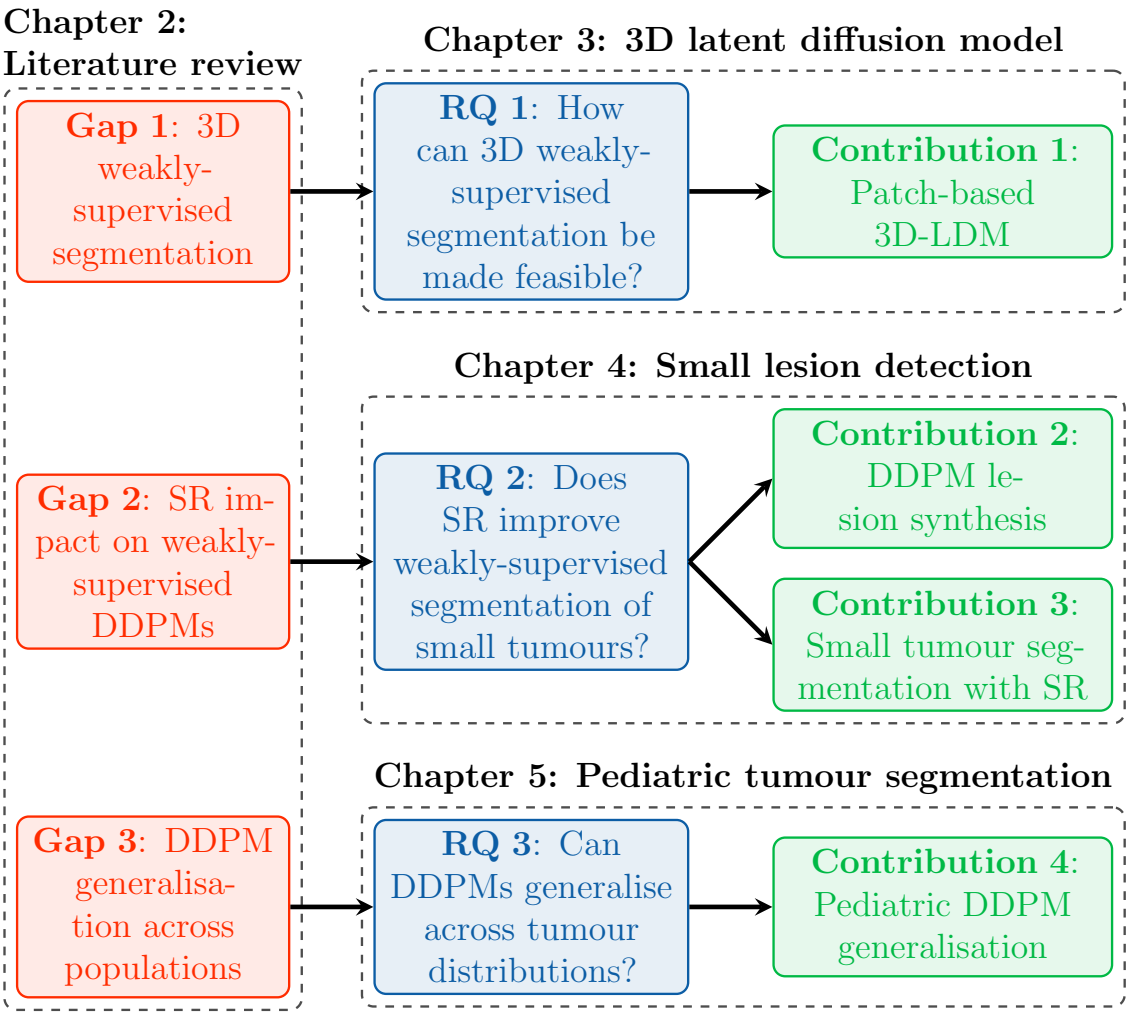


Figure 1.2: Thesis structure illustrating the identified research gaps (red), corresponding research questions (blue), and contributions of each project (green). Dotted lines with labels indicate the chapters in which each component is introduced and discussed.

Chapter 5 extends the established framework to the paediatric domain, evaluating its generalisability under substantial distributional shift and limited data availability. It examines the extent to which a model trained on adult data can be applied to a different population with distinct imaging characteristics. Particular focus is placed on the role of the encoding mechanism in supporting transferability, and how fine-tuning influences segmentation performance on the state-of-the-art dataset (see Section 5.2). The chapter also includes the analysis of the model's performance on a separate private cohort of paediatric brain tumour patients in Section 5.3 to investigate the generalisability beyond the state-of-the-art dataset. The chapter

concludes with a discussion of the limitations and a summary of the key findings in Section 5.4.

Finally, **Chapter 6** synthesises the main findings of the thesis, outlines their clinical relevance, and reflects on the broader implications for weakly-supervised medical image segmentation. It concludes by identifying limitations and proposing avenues for future research to advance the clinical applicability of weakly-supervised anomaly detection with DDPMs with a focus on paediatric brain tumour segmentation.

Notation remarks

A Note on Terminology: Throughout this dissertation, the term "resolution" is used to refer to the spatial dimensions of the image grid (e.g., 256×256 pixels), aligning with standard nomenclature in Computer Vision and DL literature. This is distinct from the clinical definition of "spatial resolution", which refers to the physical dimensions of each voxel in millimetres. Where physical voxel size is discussed, it is explicitly stated in metric units to avoid ambiguity.

Chapter 2

Literature Review

This chapter presents and critically reviews the most relevant literature on brain tumour segmentation using deep learning (DL). Section 2.1 examines the availability of data, state-of-the-art DL approaches and models, and evaluation metrics. The section concludes by addressing the challenges and limitations of state-of-the-art approaches. Section 2.2 proposes the idea and highlights the benefits of reducing supervision in medical image segmentation with a focus on brain tumours. The section introduces the concept of anomaly detection with generative models as a means to reduce the required label density during training. Anomaly detection methods utilising traditional generative models are reviewed, which transitions into the motivation behind using denoising diffusion probabilistic models (DDPMs), a novel class of generative models, in Section 2.3. The section highlights the advantages of DDPMs in medical imaging, particularly in brain tumour segmentation, and concludes by summarising the limitations and shortfalls of currently researched DDPMs used for anomaly detection in Section 2.4. This section motivates this research project and builds the foundation for the subsequent chapters.

This chapter provides the necessary context for the research project, focusing on the current state-of-the-art in brain tumour segmentation using DL. For a more detailed understanding of the introduced concepts, the interested reader is referred to the cited literature, including Goodfellow et al. (2016b) and Bishop and Bishop (2024c).

2.1 State-of-the-art brain tumour segmentation

The origin of most approaches for medical image segmentation is linked to advancements in visual recognition tasks in natural images, including classification and object detection. The success of DL in these domains can be predominantly attributed to the availability of data and the emergence of powerful computing infrastructure (Litjens et al., 2017; Ronneberger et al., 2015; Shen et al., 2017). As a result, traditional machine learning (ML) models have been gradually replaced by more advanced DL models. The latter are characterised by their ability to automatically extract essential features in contrast to the manual feature design by experts in traditional ML. Nowadays, the state-of-the-art in image processing with DL is built upon the introduction of convolutional neural networks (CNNs) and their capacity to efficiently learn hierarchical features from structured data such as images (Goodfellow et al., 2016a; He et al., 2016; Krizhevsky et al., 2012; Minaee et al., 2021).

The success of DL models for natural image analysis has substantially influenced research targeting medical imaging data. Much of the research focuses on translating knowledge and techniques to medical imaging data, which predominantly includes the adaptation to the new data domain (Litjens et al., 2017; Shen et al., 2017). This adaptation poses challenges, particularly due to differences in data characteristics, such as higher dimensionality, varying features, and diverse image compositions. For instance, in brain tumour segmentation, data typically consists of 3D magnetic resonance imaging (MRI) volumes with multiple sequences, each contributing complementary information to capture the full extent of potential lesions. Consequently, these sequences are typically combined as input to the model for a comprehensive lesion assessment, which is fundamentally different from the RGB channels in natural images (Avesta et al., 2023; Litjens et al., 2017; Singh et al., 2020; X. Zhou et al., 2018). Occasionally, models and concepts are developed specifically for medical imaging, making substantial contributions to the field of DL. A prominent example is the U-Net model by Ronneberger et al. (2015), now considered the gold standard for image segmentation across domains.

2.1.1 Data availability

A key requirement for the successful application of any DL-based method is the availability of data. Computer vision and image processing with DL was markedly influenced by the availability of open-source dataset collections, including ImageNet (Deng et al., 2009) and Microsoft COCO (Lin et al., 2014). The availability of large-scale, annotated datasets in natural image domains has enabled the development of increasingly complex DL models. These models are capable of learning more expressive feature representations and demonstrate strong generalisation to unseen data. Empirically, model performance has been shown to improve logarithmically with dataset size (Sun et al., 2017), reinforcing the importance of data availability in driving progress. In addition, the heterogeneity of the available dataset is of great importance to generalise well to unseen examples of the same distribution (He et al., 2016; Soomro et al., 2023; Varoquaux & Cheplygina, 2022; Willemink et al., 2020).

In comparison to natural images, medical imaging shows an entirely different landscape. The availability of data in this domain is substantially limited due to various reasons, including but not limited to (Kush et al., 2020; Varoquaux & Cheplygina, 2022; Willemink et al., 2020; Wirth et al., 2021):

1. ethical restrictions to share sensitive patient information,
2. deficiencies in hospital infrastructure and data collection systems hindering data sharing, and
3. the costly annotation process for high-dimensional medical imaging data.

Ethical restrictions, risks and data sharing infrastructure

Ethical considerations are essential in research involving medical data, as the data contains identifiable health information. Accordingly, studies using medical imaging must follow stringent ethical guidelines and data protection laws, with strong emphasis on safeguarding patient privacy and confidentiality. The ethical review process involves a careful assessment of potential risks and benefits to participants. An human research ethics committee (HREC) oversees this evaluation to protect participants

from harm while maximising the study's scientific and societal value (Herington et al., 2023; Kush et al., 2020; Larson et al., 2020; Willemink et al., 2020).

Typically, medical image research involves the retrospective analysis of existing patient records rather than the collection of new data. Consequently, the main risk to participants is the potential exposure of identifiable information, and a waiver of explicit consent is often granted after evaluation by the HREC. In such cases, anonymity is ensured through the removal of personal identifiers (e.g., name, date of birth, address, etc.) and information embedded in the imaging data, either in Digital Imaging and Communications in Medicine (DICOM) metadata or the image itself (Diaz et al., 2021; El Emam, 2013; J et al., 2025). Data must also be securely stored and accessible only to authorised personnel, and ethics approval is granted once these safeguards are demonstrated. The review process is often lengthy due to the many hurdles that must be addressed. Research involving paediatric data adds further complexity, as this vulnerable population requires additional safeguards and their limited capacity to provide informed consent complicates the approval process (Downie et al., 2007; Herington et al., 2023; Kush et al., 2020; Larson et al., 2020; Willemink et al., 2020).

In addition to ethical restrictions, the sharing of medical imaging data is tightly regulated (Hollis, 2016; Knoppers & Thorogood, 2017). Such data is usually stored on secure hospital servers and accessible only to institutional affiliates, such as medical practitioners. External researchers are typically denied access due to strict privacy regulations. Overcoming this limitation requires close collaborations or formal processes that affiliate researchers with the hospital, ensuring compliance with institutional ethics and data protection protocols. Although stored within hospitals, medical imaging data is rarely curated for research. Preparing suitable datasets demands considerable time, resources, and financial investment, yet most hospitals lack the budget and infrastructure to support such efforts, further constraining accessibility (Herington et al., 2023; Hollis, 2016; Knoppers & Thorogood, 2017; Willemink et al., 2020).

Relation to thesis

The ethical and regulatory considerations outlined above directly applied to the acquisition of one of the paediatric datasets used **in this thesis**. Full compliance with institutional ethics protocols was required, including de-identification and hospital affiliation, with further details provided in Section 5.3.

Annotation requirements and bias

Even if all administrative and regulatory hurdles are overcome, medical images in isolation are typically insufficient for diagnostic DL models and require additional information. The latter includes ground-truth labels for segmentation or classification. This information is crucial for effective model training as it associates visual appearance with the respective pathology, which provides guidance throughout the training process (Baid et al., 2021; Willemink et al., 2020). However, the annotation process of medical data is limited by two major factors compared to crowd-sourced or corporation-led annotation of natural images:

1. annotation of medical images is time-consuming due to the high-dimensionality of the data involving multi-sequence and complex imaging modalities such as MRI, and
2. precise annotations are challenging and require domain experts (e.g. radiologists) to ensure high-quality labels.

Both challenges are being further exacerbated in granular tasks such as segmentation, which necessitates high-quality annotations on a per-pixel level in contrast to the patient-level. The combination of these factors results in a high cost of annotation, which is often prohibitive for large-scale datasets (Baid et al., 2021; Bakas et al., 2017; Tejani et al., 2024; Willemink et al., 2020).

The requirements for data acquisition, preparation, and sharing introduce another critical issue: ***bias***. Due to the high costs and long timelines of collection and annotation, available datasets are often limited in size and diversity (Tejani et al., 2024; Willemink et al., 2020). This is evident in many state-of-the-art datasets (Baid et al., 2021; Bakas et al., 2017), which focus on a few pathologies and lack transparency

about the represented populations. The resulting ***population prevalence bias*** or ***disease prevalence bias*** can hinder performance and generalisability (Ali et al., 2024; Habib et al., 2023; Tejani et al., 2024; Willemink et al., 2020), since models depend heavily on data quality and accurate representation of pathology distributions across populations. Vulnerable groups, including children, are particularly underrepresented because of ethical concerns and associated risks. These limitations raise concerns about the real-world applicability of models trained on restricted datasets and highlight the need for diverse, large-scale, open-source collections to enable robust validation (Ali et al., 2024; Baid et al., 2021; Bakas et al., 2018; Downie et al., 2007; Habib et al., 2023; Herington et al., 2023; Soomro et al., 2023; Tejani et al., 2024; Willemink et al., 2020).

The second form of bias arises from the visual characteristics of processed imaging data, which are strongly influenced by scanner protocols, including acquisition parameters and the manufacturer of the equipment. This issue, known as ***single-source bias***, limits model generalisability, as a DL model trained on data from one institution may not perform effectively on data from another. Moreover, capturing a wide range of demographic, socioeconomic, and geographic diversity requires data sourced from multiple institutions to facilitate robust model performance (Soomro et al., 2023; Tejani et al., 2024; Willemink et al., 2020).

The third major source of bias arises from the annotation process itself. Visual assessment is inherently subjective and influenced by the annotator's expertise, knowledge, and the quality of the data. Brain tumours are especially difficult to annotate accurately due to their diffuse and often ambiguous appearance (Bi et al., 2019; Menze et al., 2015; Willemink et al., 2020). Lesions differ widely in shape, size, and location, while image artefacts can further obscure their extent. As a result, even expert annotations can be inconsistent and error-prone, introducing ***annotation bias*** that directly affects model performance and generalisability (Baid et al., 2021; Bakas et al., 2017; Menze et al., 2015; Soomro et al., 2023; Willemink et al., 2020).

Relation to thesis

This research project addresses this problem by relaxing the annotation requirements from voxel-level to volume-scale labels, as outlined in Sections 2.2 and 2.3. This reduction in supervision lowers the cost of annotation, enables the use of broader clinical datasets, and helps mitigate biases introduced through expert labelling and disease-specific data availability due to the simpler annotation process.

Public datasets: The BraTS dataset

Each limitation and the resulting bias substantially influence data quantity and quality. The scarcity of high-quality annotated data is a major hurdle for the implementation of DL approaches in clinical routine. As a result, there is a considerable effort to provide publicly available datasets to address the aforementioned limitations. The most prominent dataset in the context of brain tumour segmentation originates from the annual brain tumor segmentation (BraTS) challenge, which includes a collection of MRI scans of glioma patients and ground-truth annotations for segmentation. The dataset is widely used in the research community and has been the foundation for the development of various DL models for brain tumour segmentation (Baid et al., 2021; Bakas et al., 2017, 2018; Menze et al., 2015).

The BraTS dataset, shown in Figure 2.1, encompasses multi-sequence MRI scans of adult glioma patients, featuring standard clinical sequences. Each subject includes T_1-weighted (T_1w) volumes with and without contrast, and T_2-weighted (T_2w) imaging including T_2-fluid attenuated inversion recovery (T_2-FLAIR). These sequences, obtained from the cancer imaging archive (TCIA) (Clark et al., 2013), are co-registered to a standardised atlas of normal adult human brain structure (Rohlfing et al., 2010). The co-registration uses the T_1 contrast-enhanced (T_1ce) volume, as it provides native 3D resolution, while the other sequences are obtained through planar 2D imaging. To generate volumetric scans of these sequences, resampling is applied to achieve an isotropic voxel size of $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ (Bakas et al., 2017; Menze et al., 2015). The most recent iteration of the dataset includes 1251 annotated adult subjects

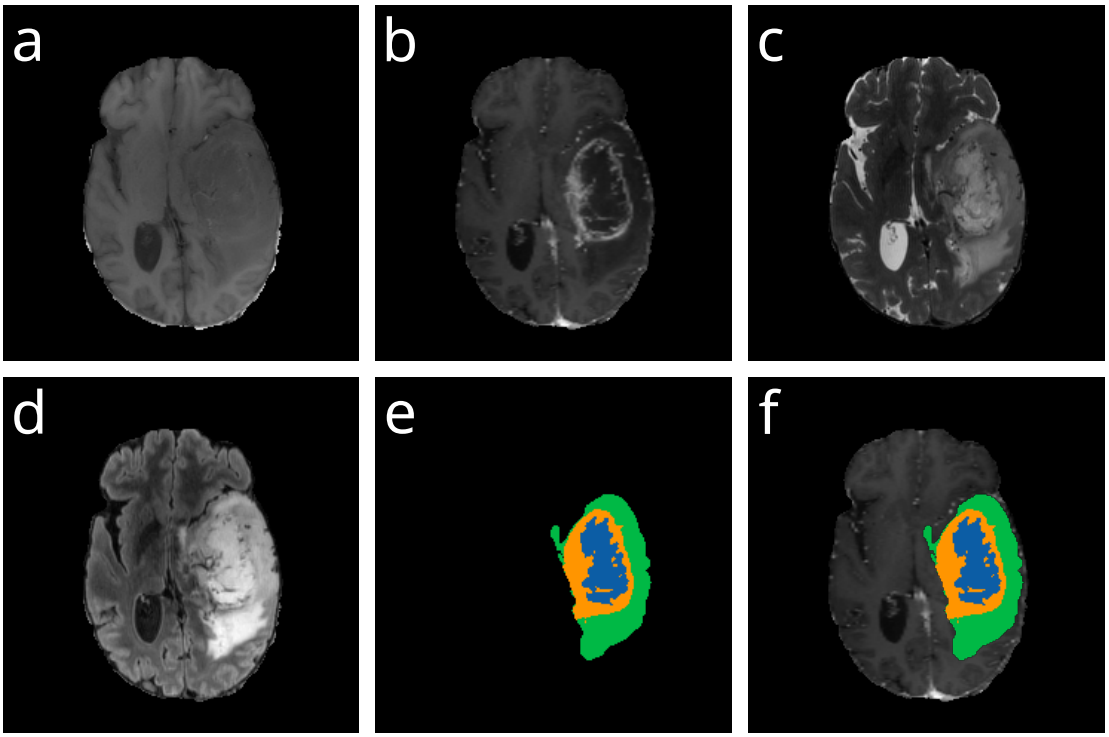


Figure 2.1: Sample from the BraTS 2023 dataset. It includes four MRI sequences: (a) T_1w, (b) T_1ce, (c) T_2w, and (d) T_2-FLAIR. Expert-provided ground-truth annotations for segmentation are shown in (e), and overlaid on the T_1ce sequence in (f). The annotations distinguish ET, NCR, and ED regions.

with low-grade glioma (LGG) and high-grade glioma (HGG). A separate validation dataset with non-public annotations is provided to evaluate the model's performance on unseen data as part of the annual challenge (Baid et al., 2021).

The dataset contains segmentation labels for three key regions: Gadolinum-enhancing tumour (ET), necrotic tumour core (NCR), and peritumoral edematous and invaded tissue (ED) (see Figure 2.1e). The ET represents areas with contrast agent leakage due to a disrupted blood-brain barrier, appearing hyper-intense in T_1ce images, indicative of high-grade lesions. The NCR marks the non-enhancing part of the tumour, typically hypo-intense in T_1ce images, corresponding to necrotic tumour areas. Finally, the ED is associated with tumours that restrict blood flow, leading to fluid accumulation in surrounding tissue, and appears hyper-intense in T_2-FLAIR imaging (Baid et al., 2021; Bakas et al., 2017).

To ensure high-quality labels, the dataset underwent a rigorous annotation process. Initially, multiple medical experts provided manual annotations. With advancements in DL segmentation, the organisers now use ensemble predictions from top-performing challenge models, which are refined by neuroradiology experts and approved by board-certified neuro-radiologists (Baid et al., 2021). This approach mitigates annotation bias and marks a step toward building a robust dataset. Furthermore, data from multiple institutions are included to reduce single-source bias (Baid et al., 2021; Bakas et al., 2017; Willemink et al., 2020).

Since its inception in 2012, the BraTS challenge has been continuously updated. These updates have included the addition of more subjects to improve model robustness, and the introduction of additional tasks including the prediction of genetic characteristic and patient survival (Baid et al., 2021; Bakas et al., 2018). In 2023, the dataset was further expanded to include a wider range of populations, such as paediatric patients (Kazerooni et al., 2024), and additional demographic groups through BraTS-Africa, which includes lower-resolution imaging data from low- and middle-income countries (Adewole et al., 2023). To ensure high-quality standards regardless of dataset scale and population, all BraTS cohorts adhere to the same rigorous, multi-expert annotation protocol detailed in the previous section. These efforts aim to reduce both population prevalence bias and disease prevalence bias, allowing to enhance the generalisability of models across diverse populations and demographics (Adewole et al., 2023; Baid et al., 2021; Kazerooni et al., 2024).

The most recent iteration of the dataset, combining data from all BraTS sub-challenges, includes 1470 fully annotated individuals with glioma (Baid et al., 2021). Due to its size and high-quality annotations, the dataset has become the gold standard for brain tumour segmentation. However, despite the extensive efforts by the organisers, the dataset remains orders of magnitude smaller than what is commonly expected for developing truly generalisable DL models (Soomro et al., 2023; Varoquaux & Cheplygina, 2022; Willemink et al., 2020).

2.1.2 Evaluation metrics

Evaluating model performance is arguably the most important step in development, as it provides insights into the model's capabilities, its limitations, and the potential for improvement. In addition, it allows the comparison to other approaches utilising a common dataset such as BraTS. This requires ground-truth annotations against which predictions can be assessed, irrespective of the level of supervision used during training (Baid et al., 2021; Bakas et al., 2017, 2018; Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018).

The evaluation of brain tumour segmentation models is performed using a combination of metrics that assess different performance aspects (Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018). Evaluation of segmentation performance typically relies on overlap-based metrics, which quantify per-voxel agreement between predicted and ground-truth labels. In contrast, boundary-distance-based metrics, which assess spatial alignment along region contours, are rarely reported in the literature. This work therefore adopts overlap-based metrics as the primary evaluation approach, which are outlined in the following section.

Overlap-based metrics

As segmentation unites classification with localisation, standard classification metrics obtained from the confusion matrix can be utilised to quantify the model's performance (Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018). In the case of brain tumour segmentation, the model is tasked to correctly classify each pixel/voxel as either tumour (positive class) or healthy tissue including the surrounding background (negative class). The positive class can be further split into the distinct tumour sub-categories for a more defined segmentation of the lesion (see Section 2.1.1). The respective metrics are then obtained using the confusion matrix (Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018).

Among the metrics derived from the confusion matrix, the ***Dice similarity coefficient (DSC)*** has emerged as the standard evaluation metric. It balances

precision and recall and is particularly sensitive to class imbalance, making it well-suited for tumour segmentation tasks where lesions are often small (Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018):

$$\mathrm{DSC} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{2.1}$$

For completeness, it is worth mentioning that the DSC is related to the Jaccard-index or intersection over union (IoU) as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} - \mathrm{TP}} = \frac{\mathrm{DSC}}{2 - \mathrm{DSC}} \tag{2.2}$$

The IoU calculates the ratio of the intersection of the predicted and ground-truth segmentation to their union. However, due to the lack of re-weighting of true positives, the IoU tends to be more penalising, especially for small lesions, when compared to the DSC (Taha & Hanbury, 2015; Yeghiazaryan & Voiculescu, 2018).

While overlap-based metrics are the de facto standard for segmentation, they are fundamentally shape-unaware and can be sensitive to lesion size variability (Maier-Hein et al., 2024). Boundary-distance metrics are often proposed as a complement; however, they were excluded from this study due to their extreme sensitivity to isolated false positives. In anomaly detection, where generative models may produce small outlier detections, distance-based measures can result in unstable metrics that do not accurately reflect clinical utility. Consequently, this research maintains a focus on overlap-based metrics to ensure robustness against outliers and to facilitate direct comparison with existing frameworks (Baur et al., 2021; Karimi & Salcudean, 2020; Sanchez et al., 2022; Wolleb et al., 2022).

Relation to thesis

In this research project, the DSC score was selected as the primary metric to ensure direct comparability with existing state-of-the-art benchmarks. While distance-based metrics offer insight into contour alignment, they were excluded due to their instability in the presence of isolated outliers. Prioritising overlap-based measures maintains methodological consistency with established anomaly detection frameworks that predominantly rely on these metrics for performance evaluation.

Generative metrics

Beyond conventional segmentation metrics, this work utilises generative metrics to assess the quality and anatomical plausibility of synthetic tumour samples. Generative metrics capture visual fidelity and distributional realism. The following section introduces the selected metrics used for evaluating synthetic data in Section 4.3.

The peak signal-to-noise ratio (PSNR) is a traditional image quality metric used to measure the fidelity of reconstructed or generated images relative to a ground-truth reference. It is defined as

$$\mathrm{PSNR(R,G)} = 10 \cdot \log_{10} \left(\frac{L^2}{\mathrm{MSE(R,G)}} \right) \quad , \tag{2.3}$$

where L denotes the dynamic range of the pixel values (e.g., 255 for 8-bit images). Higher PSNR values indicate lower reconstruction error and better signal fidelity. While widely utilised in reconstruction tasks, PSNR primarily assesses pixel wise similarity and may correlate poorly with human perception or the diagnostic clarity required in clinical practice (Horé & Ziou, 2010; Z. Wang et al., 2004). However, it remains a standard baseline for ensuring raw signal consistency before applying more complex structural evaluations.

The structural similarity index (SSIM) assesses image similarity by comparing structural information between a generated and reference image. It evaluates luminance, contrast, and structure within local image patches:

$$\text{SSIM}(R, G) = \frac{(2\mu_R\mu_G + c_1)(2\sigma_{RG} + c_2)}{(\mu_R^2 + \mu_G^2 + c_1)(\sigma_R^2 + \sigma_G^2 + c_2)} \quad , \tag{2.4}$$

where R and G are the real and synthetically generated images, respectively, μ_R and μ_G are their local means, σ_R^2 and σ_G^2 are their local variances, and σ_{RG} is the local cross-covariance. The constants c_1 and c_2 are stability constants to avoid division by zero. SSIM emphasises perceptual quality by focusing on structural similarity rather than raw pixel fidelity, making it more aligned with human visual perception than traditional metrics like PSNR (Z. Wang et al., 2003).

The multi-scale structural similarity index (MS-SSIM) extends the SSIM metric by evaluating image similarity across multiple scales, capturing both fine and coarse structural details. It aggregates SSIM scores at different resolutions using a weighted geometric mean:

$$\text{MS-SSIM}(R, G) = \prod_{j=1}^{M} [\text{SSIM}_j(R, G)]^{\alpha_j} \tag{2.5}$$

M describes the number of scales, $\text{SSIM}_j(R, G)$ is the SSIM score at scale j, and α_j are weights that control the contribution of each scale. This multi scale approach allows MS-SSIM to better reflect perceptual consistency across varying spatial frequencies by incorporating luminance, contrast, and texture (Z. Wang et al., 2003). These structural metrics offer improved alignment with radiological interpretation as they focus on the preservation of tissue boundaries and anatomical features rather than raw pixel values (Renieblas et al., 2017; L. Zhang et al., 2011). By assessing structural integrity at multiple resolutions, these metrics capture both fine and coarse details critical for identifying pathological anomalies.

The Fréchet inception distance (FID) is a widely used metric for evaluating generative models. It compares the distributions of feature representations between real (R) and generated (G) images using a pre-trained Inception network. In contrast, the inception score (IS) considers only G in its calculation. The FID computes the Fréchet distance between the multivariate Gaussian distributions of R and G

$$\text{FID}(R, G) = |\boldsymbol{\mu}_R - \boldsymbol{\mu}_G|_2^2 + \text{Tr}\left(\boldsymbol{\Sigma}_R + \boldsymbol{\Sigma}_G - 2(\boldsymbol{\Sigma}_R\boldsymbol{\Sigma}_G)^{\frac{1}{2}}\right) \quad , \tag{2.6}$$

with both distributions defined by their mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ (Dowson & Landau, 1982; Heusel et al., 2017). Low FID scores indicate that the feature distribution of generated images closely matches that of real images, as evaluated by a pre-trained Inception network. A low score reflects both high visual quality and diversity, showing that the model produces realistic and varied outputs that are perceptually indistinguishable from real images in texture, structure, and detail (Heusel et al., 2017).

Similarly, the learned perceptual image patch similarity (LPIPS) metric measures perceptual similarity between two images using deep features extracted from pre-trained networks such as AlexNet or VGG. For real images R and generated images G, their activations at multiple layers l are compared as follows:

$$\text{LPIPS}(R, G) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (f_l(R) - f_l(G))\|_2^2 \tag{2.7}$$

where $f_l(R)$ are the features at layer l, w_l are learned per-channel weights, and H_l, W_l are the spatial dimensions. The $\odot$ operator denotes element-wise multiplication. LPIPS captures high-level perceptual differences by comparing deep semantic representations rather than raw pixels, aligning well with human visual judgement (R. Zhang et al., 2018). While FID and LPIPS bridge the gap between pixel fidelity and clinical relevance, their effectiveness depends on the domain of the feature extractor, as traditional networks trained on natural images may not fully capture the nuanced anatomical details of medical MRI (Barratt & Sharma, 2018; Bercea et al., 2024).

The Fréchet autoencoder distance (FAD) was proposed as a domain-specific variant of FID for medical imaging. It computes the Fréchet distance between autoencoder (AE)-derived feature representations of real and generated images. Unlike FID, which uses Inception-v3 trained on natural images, FAD employs a domain-specific AE pretrained on brain MRI, ensuring the extracted features reflect medically relevant anatomical content (Buzuti & Thomaz, 2023). This adaptation improves alignment with expert assessment in neuroimaging, enhancing the metric's relevance for evaluating anatomical plausibility in generative medical models.

2.1.3 Key concepts and network architectures

With its ability to automatically extract relevant features, DL has transformed medical image analysis and enabled more accurate diagnostic tools. CNNs are specifically designed to detect spatial hierarchies in grid-like data (e.g. images and volumes). By capturing local patterns and ensuring translation equivariance, they provide a strong foundation for tasks such as tumour detection (Goodfellow et al., 2016a; He et al., 2016; Krizhevsky et al., 2012; Ronneberger et al., 2015). Convolutional layers have largely replaced the fully connected layers of early DL models because they handle spatial information more effectively and are computationally more efficient. To address limitations in training deeper networks, methods such as residual connections and attention mechanisms were introduced, which further improve accuracy and efficiency in medical image analysis. These advances are now embedded in specialised architectures for brain tumour segmentation (Bishop & Bishop, 2024a, 2024b; Goodfellow et al., 2016a; He et al., 2016; Krizhevsky et al., 2012; H. Yu et al., 2021). The key principles will be outlined in the following section.

Convolutional neural networks

CNNs are foundational to modern image analysis frameworks due to their ability to learn spatial hierarchies of features from raw image data. Their architecture leverages key design advantages (Bishop & Bishop, 2024a, 2024b; Goodfellow et al., 2016a; He et al., 2016; Krizhevsky et al., 2012; H. Yu et al., 2021): (1) Compatibility with varying input sizes, (2) sparse connections for reduced computational cost, and (3) translation equivariance, ensuring consistent feature detection across spatial locations.. Given these advantages, CNNs have become the standard architecture for image analysis tasks, including brain tumour segmentation. Their ability to capture local patterns and spatial hierarchies makes them particularly effective for identifying and delineating tumours in medical images (Litjens et al., 2017; R. Wang et al., 2022; H. Yu et al., 2021). As a result, CNNs form the backbone of the segmentation frameworks introduced and used throughout this research project.

Attention mechanism

The second important foundational building block in DL for image analysis is the attention mechanism, originally introduced for sequence modelling by Vaswani et al. (2017). In contrast to convolutions, which operate with a limited receptive field, attention enables global feature aggregation by dynamically weighting relationships across the entire input. This makes it particularly valuable for segmentation tasks, where contextual dependencies and long-range interactions are crucial (Minaee et al., 2021; R. Wang et al., 2022).

At its core, the attention mechanism computes three projection matrices from the input features: the query matrix $\boldsymbol{Q} \in \mathbb{R}^{N \times D_\text{h}}$, the key matrix $\boldsymbol{K} \in \mathbb{R}^{N \times D_\text{h}}$, and the value matrix $\boldsymbol{V} \in \mathbb{R}^{N \times D_\text{out}}$. Here, N denotes the number of spatial positions, D_h is the hidden or embedding dimension for attention, and D_out defines the dimensionality of the output representation. These are used in the ***self-attention*** operation to weigh the contribution of each element based on its relevance to all others:

$$\boldsymbol{Y} = \text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left[\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{D_k}}\right]\boldsymbol{V} \tag{2.8}$$

The Softmax ensures the attention weights are normalised and the scaling term $\sqrt{D_k}$ preserves gradient stability (Bishop & Bishop, 2024e; Vaswani et al., 2017). To capture different relational patterns in parallel, multi-head attention computes several attention maps using independent sets of weights, combining them into a richer representation through linear projection (Vaswani et al., 2017).

Relation to thesis

Given its capacity to model global dependencies, attention mechanisms have been widely adopted in image analysis tasks. They constitute the architectural backbone of the proposed framework described in Section 2.3 and utilised **throughout this thesis**. In addition, they also facilitate conditional generation, as discussed in Section 2.3.3.

Network architectures for semantic segmentation

Segmentation has progressed beyond traditional image classification by requiring dense, pixel-wise predictions that preserve spatial information. Classification networks use fully connected layers that collapse spatial context into a single vector, making them unsuitable for segmentation. To overcome this, Long et al. (2015) proposed fully convolutional networks, which replace fully connected layers with convolutional ones. This allows spatially-aware, dense predictions across the image and enables the generation of segmentation masks through upsampling with either interpolation or transposed convolutions.

This concept was extended by Ronneberger et al. (2015) with the introduction of the ***U-Net***, which combines a contracting path for context with an expansive path for localisation. Skip connections between corresponding layers restore spatial detail lost during downsampling, enabling precise pixel-level predictions. The U-Net has since become the standard architecture in medical image segmentation, and has been adapted to handle 3D data (Çiçek et al., 2016) and integrate advanced methods to enhance performance (Dhariwal & Nichol, 2021; Isensee et al., 2021; Rombach et al., 2022; Ronneberger et al., 2015).

Relation to thesis

The ability of U-Net to recover spatial resolution through its expansive path makes it particularly well-suited as the backbone architecture for the model framework introduced in Section 2.3, which serves as the foundational architecture **across all chapters of this thesis**.

2.1.4 Supervised learning: the gold standard

Brain tumour segmentation methods can be categorised by the amount and granularity of ground-truth label information used during training. ***Supervised learning*** is widely regarded as the gold standard for this task (Goodfellow et al., 2016c; Litjens et al., 2017; R. Wang et al., 2022). Ground-truth labels provide explicit guidance that steers the model towards an optimal solution. These labels reduce uncertainty

and allow direct, task-specific optimisation. Such properties are especially important in medical imaging, where accuracy and reliability are critical.

As supervised learning relies on pixel-level annotations, the BraTS dataset is commonly used for brain tumour segmentation due to its detailed ground-truth labels. In addition, the dataset provides a comparatively large number of subjects and complimentary multi-sequence MRI scans (see Section 2.1.1). Most recent approaches incorporate the MRI sequences as a separate channel in the input layer (Boutry et al., 2020; Rosas González et al., 2020). Additionally, 3D convolutions are preferred over 2D for spatial consistency, with patch-based training often used to manage memory constraints while preserving global context (Amian & Soltaninejad, 2020; Z. Jiang et al., 2020; Y. Zhang et al., 2021).

Most recent approaches make either minor modifications to conventional architectures or combine multiple state-of-the-art techniques to enhance segmentation performance. A large body of work builds on the U-Net backbone, incorporating auxiliary DL principles such as attention mechanisms, multi-scale features, or both (W. Chen et al., 2022; Y. Feng et al., 2024; Hu et al., 2023; J. Wang, Gao, et al., 2021; Z. Zhu et al., 2023). Performance improvements have also been pursued through ensemble learning, where multiple models are aggregated using majority or weighted voting to reduce variance and improve generalisation (X. Feng et al., 2020; Y. Zhang et al., 2021). Multitask learning introduces auxiliary objectives either through composite loss functions (Agravat & Raval, 2020; X. Chen et al., 2019; Jadon, 2020; Lin et al., 2020; Salehi et al., 2017) or by incorporating shared-encoder, multi-branch architectures to provide regularisation and capture complementary information (Myronenko, 2019; Xue et al., 2018; Yi et al., 2019). Cascaded models further extend this strategy by refining predictions across sequential stages or structuring tasks hierarchically, often using curriculum learning or coarse-to-fine pipelines (Amian & Soltaninejad, 2020; Bangalore Yogananda et al., 2020; Guohua et al., 2020; Z. Jiang et al., 2020; C. Zhou et al., 2020).

Despite the diversity of architectural innovations, recent analyses suggest that model configuration is not the primary determinant of performance. Isensee et al.

(2021) and Litjens et al. (2017) both highlight that models using identical architectures can achieve widely varying results, underscoring the limited impact of structural complexity alone. In fact, Isensee et al. (2021) observe that simple architectures, when properly configured, can match or outperform more elaborate designs, suggesting a saturation in achievable performance for this task.

Relation to thesis

This thesis builds on these findings by adopting a standard U-Net architecture, closely aligned with the 2D variant of the nnU-Net framework, as the supervised baseline (see Section 4.4 and Chapter 5).

2.1.5 From convolution to Transformers: Evolving architectures in medical imaging

Rather than refining increasingly complex CNN-based networks, recent work explores transformer-based architectures for medical image segmentation. Originally developed for natural language processing, transformers offer strong global context modelling and long-range dependency capture, properties that are well-suited to tasks like brain tumour segmentation (Dosovitskiy et al., 2021; Hatamizadeh, Nath, et al., 2022; Liu et al., 2021; Vaswani et al., 2017; W. Wang et al., 2021; Z. Zhu et al., 2023). Unlike traditional CNNs, which may incorporate attention blocks to enhance feature learning, transformer-based models substitute large portions of the architecture with attention-based layers. However, replacing CNN layers with attention blocks incurs two main challenges: (1) the reduction in generalisability due to the loss of inductive bias, including translation equivariance and locality, and (2) the increased computational cost due to the quadratic complexity of the attention mechanism. Dosovitskiy et al. (2021) demonstrated to mitigate the first limitation by pre-training the model on large-scale image datasets, surpassing state-of-the-art CNNs in image classification tasks. The second limitation is addressed by utilising a patch-based input representation, where the input image is divided into patches of size $P \times P$ and embedded into a feature representation. To mitigate the high-computational

demands for large image sizes, Liu et al. (2021) introduced the Swin Transformer, which introduces a hierarchical structure with shifted windows to efficiently model local and global dependencies.

Extensions of these models adapt them for segmentation: Hatamizadeh, Tang, et al. (2022) integrate a CNN-based decoder into the Swin Transformer, W. Wang et al. (2021) combine a U-Net-like CNN with a transformer bottleneck for enhanced spatial-depth feature extraction, and Z. Zhu et al. (2023) augment Swin with an edge detection module to improve boundary delineation. A 3D adaptation of Swin has also shown improved volumetric performance by leveraging spatial depth (Liang et al., 2022). The fusion of transformers and CNNs is emerging as a key trend, aiming to balance local detail and global context for robust segmentation (Hatamizadeh, Tang, et al., 2022; Liang et al., 2022; W. Wang et al., 2021; Z. Zhu et al., 2023).

While transformer-based models have shown promising results, their performance is contingent on large-scale pre-training, which is not feasible in most medical imaging domains due to limited data availability (see Section 2.1.1). Moreover, the high computational demands of attention-based operations, particularly their quadratic complexity, pose a major barrier to clinical deployment where computational resources may be constrained.

Relation to thesis

For these reasons, **this thesis** adopts more computationally efficient CNN-based architectures, which remain more practical under real-world conditions and are well-suited to the scale of the available data.

2.1.6 Small lesion detection

Accurate segmentation models are essential not only for delineating tumour boundaries during radiologic imaging analysis but also for reliably detecting small or subtle lesions, which is particularly critical given the high clinical impact of early diagnosis. Smaller lesions are typically indicative of earlier disease stages due to the progressive nature of brain tumours. Timely diagnosis strongly influences treatment outcomes

and long-term prognosis, as therapeutic options diminish with lesion growth (Alemany et al., 2021; E. L. Chang et al., 2003; Deike-Hofmann et al., 2018; Fry et al., 2014; Goldman et al., 2017; Sadighi et al., 2018; S. Wu et al., 2021; Yamada et al., 2020). Early detection is especially critical in paediatric populations, where the developing brain is highly vulnerable to treatment-induced neurotoxicity during periods of rapid growth and maturation (de Ruiter et al., 2013; Palmer et al., 2001; Ris et al., 2001). Delayed diagnosis has been linked to persistent physical, sensory, cognitive, and neurological deficits, including seizures, hearing loss, and visual impairment. More than 60% of childhood cancer survivors experience long-term disabilities, with marked consequences for intellectual development and overall quality of life (Armstrong, 2010; de Ruiter et al., 2013; Goldman et al., 2017; Lassaletta et al., 2015; Sadighi et al., 2018; Wilne et al., 2010).

Relation to thesis

Early detection is **in this thesis** synonymous with the identification of small lesions, as tumours arise microscopically and expand over time. The earliest detectable manifestations are therefore typically small, focal abnormalities, which requires high-sensitivity imaging techniques to identify them reliably.

However, conventional MRI faces substantial challenges in identifying small and subtle lesions due to several factors, including but not limited to:

- **Resolution constraints:** Standard-resolution MRI may fail to capture fine structural details, increasing the risk of overlooking small lesions. This issue is particularly pronounced for lesions below the resolution threshold of the imaging modality, which may be obscured or misinterpreted as normal tissue (Bruno et al., 2015; Deike-Hofmann et al., 2018; Goldman et al., 2017; Lee et al., 2013; L. Zhang, Wen, et al., 2023). Additionally, MRI is susceptible to various imaging artefacts, including motion-related distortions and scanner-induced noise (Aja-Fernández & Vegas-Sánchez-Ferrero, 2016; Dale et al., 2015; Shaw et al., 2019; Sled et al., 1998), which can further obscure small lesions and complicate accurate diagnosis

- **Radiologic error:** Small or subtle lesions are among the most commonly overlooked findings in radiological assessment, often due to factors such as atypical presentation, limited visibility, or perceptual challenges inherent to the constraints of MRI (Lee et al., 2013; L. Zhang, Wen, et al., 2023).

Consequently, improving the sensitivity of imaging modalities to detect small and subtle lesions remains a key objective in neuro-oncological diagnostics. Strategies to mitigate these challenges include increasing the MRI scanner's field strength beyond 3 T and integrating specialised sequences with improved coil technology (Burkett et al., 2021; Mamlouk et al., 2017; Thust et al., 2018). Furthermore, high-resolution (HR) 3D sequences improve through-plane resolution, reducing the risk of omitting small contrast-enhancing regions due to partial volume effects (Kwak et al., 2015).

However, high-field-strength MRI remains largely restricted to specialised research facilities and presents unresolved challenges, including increased tissue heating, artefacts from field inhomogeneity, and patient discomfort, such as vertigo (Burkett et al., 2021; Kabasawa, 2021). Additionally, current 3D sequences may be suboptimal for detecting small lesions due to limited T_1w-dependent signal enhancement compared with spin-echo sequences (Thust et al., 2018), highlighting the need for alternative approaches to enhance small lesion detection in clinical practice.

What is considered a small lesion?

While the term "small lesion" has been used throughout this section to refer to lesions of limited size, its precise definition remains somewhat ambiguous. What exactly qualifies as "small" can vary depending on the clinical or technical context. Existing literature provides some guidance on thresholds for lesion size classification. W. J. Chung et al. (2012) defines a lesion as small if its diameter is less than 50 mm, whereas clinical trials commonly set the threshold for a measurable lesion at 10 mm (Henson et al., 2008) - a threshold also confirmed by Kwak et al. (2015). In contrast, studies in multiple sclerosis report lesions as small as 3 mm (Grahl et al., 2019), with the resolution of the MRI field strength and imaging protocol ultimately determining the smallest detectable lesion. Furthermore, a study on brain metastases from

melanoma, involving 224 patients, observed a median lesion diameter of 5 mm, which was associated with early diagnosis (Deike-Hofmann et al., 2018).

Relation to thesis

For the purposes of **this research project**, lesions with a diameter of less than 10 mm will be considered small, aligning with the commonly accepted threshold for measurable lesions in clinical settings (Henson et al., 2008; Kwak et al., 2015) and the Response Evaluation Criteria in Solid Tumors (RECIST) guidelines (Eisenhauer et al., 2009). The latter explicitly defines as lesions with a diameter of less than 10 mm as non-measurable and therefore not suitable for clinical assessment, particularly in the context of treatment response assessment.

While lesion diameter is frequently cited as the principal metric for assessing tumour size, its precise definition remains ambiguous across clinical guidelines. According to the RECIST criteria, diameter is defined as the longest axis measurable along the lesion's perimeter in a single plane (Dempsey et al., 2005; Eisenhauer et al., 2009). Importantly, the longest diameter should be measured in the plane in which the imaging data were acquired. For MRI, which is typically acquired in the axial plane, this implies that lesion measurements must be taken in the axial view to maintain consistency with RECIST guidelines (Eisenhauer et al., 2009). This is largely consistent with the 2D Macdonald criteria, which similarly identify the longest axial diameter but extend the measurement to a surrogate area by multiplying it with the orthogonal perpendicular diameter (Henson et al., 2008; Macdonald et al., 1990). These variations in estimation underscore the lack of standardisation in defining tumour diameter, particularly as volumetric assessments gain prominence with the widespread adoption of 3D imaging protocols (Henson et al., 2008).

Deep Learning and the detection of small lesions

Addressing the challenges of small lesion detection requires methods that overcome spatial resolution limitations while supporting radiologic interpretation. Although DL models have shown strong performance in segmenting large, well-defined lesions,

their sensitivity to lesion size and ability to detect subtle abnormalities remain underexplored. Most segmentation studies do not systematically quantify detection performance for small lesions, with only a few specialised publications focusing on this critical issue (Bria et al., 2020; Sørensen et al., 2023; S. Wu et al., 2021). This gap is particularly relevant in clinical contexts, where diagnostic accuracy depends on the reliable identification of early-stage or subtle findings.

Performance degradation with decreasing lesion size is consistently observed (L. Li et al., 2021; Sørensen et al., 2023; S. Wu et al., 2021; Yoo et al., 2021). Contributing factors include class imbalance, where tumour voxels form a small minority relative to background, compromising CNN training (Bria et al., 2020). Two key strategies are commonly adopted to address these limitations. The first involves adapting the loss function to improve sensitivity to small structures by balancing the relative contribution of lesion to background during training (Abraham & Khan, 2019; Lin et al., 2020; Y. Zhang et al., 2024).

The second strategy focuses on modifying the receptive field through multi-scale feature extraction or prediction. This includes attention-based mechanisms, inter-slice context modelling, and spatial upsampling to better preserve fine-grained details (Savelli et al., 2020; Tao et al., 2019; B. Xu et al., 2018). Some works explore ensemble models that separate small and large lesion segmentation tasks (Erdur et al., 2024), while others incorporate scale-awareness using dynamic receptive fields or dilated convolutions (Luo et al., 2024). A complementary strategy involves increasing resolution to improve sensitivity to small lesions (Wong et al., 2022).

Relation to thesis

This thesis investigates the effect of higher spatial resolution on small lesion detection performance in the context of weakly-supervised brain tumour segmentation (see Section 2.3.6)

2.1.7 Deep Learning for paediatric brain tumour segmentation

The majority of DL-based brain tumour segmentation approaches primarily focus on adult populations due to the existence of the BraTS dataset. However, the anatomical differences between paediatric and adult brains pose unique challenges that require tailored solutions. These differences extend not only to the tumours themselves but also to the surrounding healthy tissue, which undergoes continuous development from neonatal stages to adulthood (***Central Nervous System Tumours***, 2021; Drai et al., 2022; Kazerooni et al., 2024; Shaari et al., 2021). The scarcity of a comprehensive paediatric dataset further complicates this task, as most studies in the field are based on small sample sizes, often in the low double digits (J. Huang et al., 2022; Shaari et al., 2021). As a result, DL methods typically adapt adult-trained architectures. Recent studies report modest gains from cascaded or ensemble nnU-Net variants but persistent failures to reliably delineate cystic tumour components (Bengtsson et al., 2025; Kazerooni et al., 2024; Mulvany et al., 2024).

2.1.8 Limitations of state-of-the-art supervised Deep Learning

The limitations of state-of-the-art supervised DL approaches for brain tumour segmentation can be summarised as follows:

1. **Insufficient annotated data**: High annotation costs and limited dataset availability constrain training diversity, reducing generalisability across patient populations (e.g., adult to paediatric) and tumour types.
2. **Limited sensitivity to small lesions**: Supervised models underperform on small or early-stage tumours due to class imbalance and lack of dedicated datasets, impeding detection in clinically critical scenarios.
3. **Stagnation in architectural innovation**: Increasing model complexity yields diminishing returns, highlighting the need for alternative strategies beyond conventional supervised learning to improve clinical relevance and adaptability.

Firstly, the scarcity of annotated data remains a major barrier to the clinical integration of supervised DL models. Label acquisition in medicine is particularly burdensome, requiring time-intensive, specialised effort from clinical experts (J. Peng et al., 2020; Varoquaux & Cheplygina, 2022; D. Wang et al., 2020; Willemink et al., 2020; Y. Zhou et al., 2019). Consequently, medical imaging datasets are often orders of magnitude smaller than those in other domains that have driven progress in ML, limiting the diversity and scale needed for effective generalisation (Cordts et al., 2016; Deng et al., 2009; Fink et al., 2020).

Beyond quantity, annotation quality introduces further challenges. Labelling bias, inter-observer variability, and imaging ambiguities such as partial volume effects limit the reliability of voxel-wise ground truth (Baid et al., 2021; H. Jiang & Nachum, 2019; Mehta et al., 2021; Varoquaux & Cheplygina, 2022). These factors constrain the performance ceiling of supervised models, which remain bound by the fidelity of their training data. While efforts such as multi-institutional datasets and annotation standardisation seek to reduce bias (Baid et al., 2021; Bakas et al., 2017; Bernhardt et al., 2022), it is unclear whether reported gains reflect real-world generalisability or increased overfitting to flawed benchmarks (Müller et al., 2020).

Limited data availability directly constrains model generalisability across both patient populations and tumour types. Brain tumour segmentation research has mainly focused on adult gliomas, largely due to the availability of well-annotated public datasets (Baid et al., 2021; Bakas et al., 2017, 2018; Kazerooni et al., 2024). In contrast, paediatric datasets are rare and often suffer from small sample sizes or narrow age coverage, which is problematic given the substantial developmental changes in the brain from infancy to adolescence (Drai et al., 2022; Shaari et al., 2021). These developmental and molecular differences raise important concerns about the applicability of adult-trained models to paediatric cases (d'Amati et al., 2024; Drai et al., 2022; Pfister et al., 2022; Willemink et al., 2020).

A similar issue arises with tumour type diversity. Most datasets are skewed towards gliomas, particularly HGGs with diffuse boundaries. Although current models achieve high accuracy in delineating these tumours, their clinical utility

is limited when effective treatment options are absent (see Section 1.1.1) (Forst et al., 2014; Varoquaux & Cheplygina, 2022; Weller et al., 2015). The narrow focus of supervised models may also fail to capture the morphological diversity of other tumour subtypes, many of which could benefit more directly from accurate segmentation for surgical or radiotherapeutic planning. These limitations highlight the need for generalisable and flexible approaches that address the heterogeneity of both patient populations and tumour biology.

Secondly, existing studies consistently report substantial performance degradation when segmenting small brain tumours, which remain more difficult to detect than larger lesions (L. Li et al., 2021; S. Wu et al., 2021; Yoo et al., 2021). Even with explicit supervision, conventional models fail to reliably identify small lesions, reflecting limitations in discriminative feature extraction and class imbalance handling. While related work has targeted small lesion detection in multiple sclerosis, haemorrhage, and stroke (An et al., 2023; L. Li et al., 2021; Nair et al., 2020; Wong et al., 2022), focused research on early-stage brain tumours remains scarce. This is largely due to the absence of dedicated datasets capturing small glioma lesions, which restricts systematic evaluation and method development in this critical regime.

Lastly, beyond the limitations imposed by data availability and annotation quality, recent findings point to a broader stagnation in architectural innovation. Despite the introduction of increasingly complex model designs, performance gains have plateaued, with studies showing that well-configured simple architectures can rival more elaborate ones (Isensee et al., 2021; Litjens et al., 2017). This saturation suggests diminishing returns from conventional supervised learning pipelines. To advance DL-based brain tumour segmentation, future research may benefit from alternative strategies that reduce reliance on dense supervision and annotated datasets, enabling more flexible and generalisable approaches that can adapt to diverse clinical scenarios.

2.2 Advancing beyond supervised learning: Techniques to reduce label dependency

The challenge of reducing annotation requirements to train DL-based models emerges naturally from the limitations of supervised learning approaches outlined in the previous section (see Section 2.1.8). Although supervised learning remains the gold-standard, its heavy dependence on limited and expensive annotated datasets motivates the exploration of more resource-efficient alternatives. Section 2.2.1 reviews recent advances in minimising annotated data requirements with alternative learning strategies, including self-supervised, semi-supervised and transfer learning. Section 2.2.2 then introduces the concept of anomaly detection as a promising alternative to supervised learning, which can be applied to segmentation tasks without the need for dense annotations. The section introduces state-of-the-art models for this task and motivates the usage of an alternative class of generative models in the subsequent section (see Section 2.3).

2.2.1 Minimising annotated data requirements in segmentation models

One promising approach to reduce supervision requirements is the utilisation of semi-supervised and self-supervised learning. Both aim to leverage labelled and unlabelled data to improve model performance, particularly when labelled data is scarce. The idea is to learn general features from unlabelled data and fine-tune on smaller, annotated datasets (Cheplygina et al., 2019; Dosovitskiy et al., 2021; Liang et al., 2022; J. Peng & Wang, 2021; Z. Zhu et al., 2023). In contrast, transfer learning involves fine-tuning the model on a dataset different to the pre-training stage. It aims to leverage learned representations to improve performance on a related, target task with limited labelled data (Valverde et al., 2021; Zhuang et al., 2021).

Semi-supervised learning

Semi-supervised learning can be categorised into two major approaches: ***pseudo-labelling*** and ***consistency regularisation***. Pseudo-labelling generates artificial labels for unlabelled data based on model predictions, guiding further training. Consistency regularisation encourages the model to produce consistent predictions for unlabelled data under different perturbations, promoting robust learning (Chaitanya et al., 2023; Fang & Li, 2020; Thompson et al., 2022; Z. Xu et al., 2023).

Semi-supervised segmentation approaches often rely on pseudo-labelling, yet often neglect label quality, which can impair performance. To address this, techniques such as superpixel-based refinement (Thompson et al., 2022) and contrastive representation learning (Chaitanya et al., 2023) have been introduced. These methods improve pseudo-label reliability or bypass them by enforcing semantic consistency across labelled and unlabelled data. Similarly, consistency learning strategies promote prediction stability under perturbations, typically combining supervised loss with entropy-based regularisation or adversarial learning (Fang & Li, 2020).

More recent approaches incorporate bidirectional information flow between labelled and unlabelled domains. Cyclic consistency and self-ensembling teacher models improve generalisability by propagating supervision across both domains (Z. Xu et al., 2022, 2023). Others leverage uncertainty estimates to guide mutual consistency between different segmentation branches, improving the reliability of unlabelled data supervision (Y. Zhang et al., 2023).

Self-supervised pre-training

Self-supervised learning derives supervisory signals directly from data by solving pretext tasks, enabling representation learning without labels. Pre-training in this paradigm yields generalisable feature embeddings that capture salient structures, which can subsequently initialise downstream supervised models and improve performance in label-scarce settings (Dosovitskiy et al., 2021; Liang et al., 2022; VanBerlo et al., 2024; Z. Zhu et al., 2023).

Self-supervised learning strategies in medical imaging often rely on context-based reconstruction tasks to enhance feature learning. Methods such as context restoration through patch swapping (L. Chen et al., 2019), hole-filling in homogeneous supervoxel regions (Kayal et al., 2020), and masked patch reconstruction using AE-based architectures (Liang et al., 2022) aim to capture spatial dependencies and structural relationships critical for downstream segmentation tasks.

Other approaches focus on architectural adaptations rather than explicit pretext tasks. For instance, Z. Zhu et al. (2023) introduce a shifted patch tokenisation strategy to enhance the spatial inductive bias of Swin Transformers, improving segmentation accuracy through better edge representation and feature aggregation.

Transfer learning to harness domain knowledge

Transfer learning leverages pre-trained models from label-rich source domains to initialise learning in target tasks with limited annotations. By transferring learned representations, it enables efficient fine-tuning on small datasets while reducing reliance on extensive labelling and maintaining strong performance (Ardalan & Subbian, 2022; Tan et al., 2018; Zhuang et al., 2021). A crucial concept underpinning these methods is the dissimilarity between the domains; the larger the difference, the less information can be transferred, which can result in reduced training efficacy (negative transfer). Consequently, most approaches rely on the principle of domain adaptation, which aims to minimise the divergence between the source and target domains to compensate for the scarcity of data in the target domain (Ardalan & Subbian, 2022; Nalepa et al., 2019; Tan et al., 2018; Zhuang et al., 2021).

Supervised domain adaptation leverages labelled datasets (e.g. ImageNet (Deng et al., 2009)), to pre-train segmentation networks for medical imaging tasks. While this reduces the need for extensive annotations in the target domain, the disconnect between natural image datasets and the medical domain limits its effectiveness (AlAmir & AlGhamdi, 2022; Ardalan & Subbian, 2022; Maqsood et al., 2019; Tan et al., 2018; Valverde et al., 2021; Wacker et al., 2021; J. Wang, Wei, et al., 2021). Kaur et al. (2019) show that fine-tuning all layers of pre-trained models outperforms

selective fine-tuning of deeper layers, even when the latter is considered state-of-the-art. Ghaffari et al. (2022) apply a model trained on pre-operative BraTS data to a small post-operative dataset with only 15 subjects. Despite substantial differences in tumour appearance due to surgical intervention, strong segmentation performance is achieved, aided by extensive data augmentation.

Unsupervised domain adaptation aims to adapt models to new target domains without requiring labelled data, by leveraging knowledge from a related, annotated source domain. Tokuoka et al. (2019) use a Cycle-generative adversarial network (GAN)-based approach to translate labelled brain tissue images to unlabelled brain tumour images, incorporating a segmentation-aware discriminator to maintain label consistency during adaptation. Similarly, Dong et al. (2020) transfer glioma images into a source domain with general tumour labels, iteratively refining the model using both ground-truth and pseudo-labelled data.

Limitations of learning with reduced annotations

Approaches that reduce annotation requirements, such as self-supervised, semi-supervised, and transfer learning, have demonstrated substantial potential in addressing the challenges arising from data scarcity in medical image segmentation. These methods excel in leveraging large amounts of unlabelled data or previously acquired knowledge, leading to improved generalisation, robustness, and reduced reliance on detailed annotations. As such, they represent a major step forward in mitigating the limitations of fully supervised learning.

However, despite these considerable benefits, these approach still rely fundamentally on detailed annotations of a similar task and are therefore constrained by the same limitations of supervised methods (see Section 2.1.8). Furthermore, transfer learning in medical imaging faces challenges, including the need to adapt pretrained models to volumetric data, determining which layers to fine-tune, and limitations in domain adaptation techniques, which fail to generate adequate geometric and distributional adaptations for target domains (Ardalan & Subbian, 2022; Choong & Hameed, 2021; Kaur et al., 2019; Q. Li et al., 2020; Wacker et al., 2021; J.-Y. Zhu et al., 2020).

2.2.2 The concept of anomaly detection

As the limitations of relying on detailed annotations for brain tumour segmentation become increasingly evident, particularly in terms of cost, subjectivity, and scalability, a central question emerges:

> ***Can we build models that no longer depend on fine-grained, expert-level annotations?***

A positive answer to this question would mark a fundamental shift in medical image analysis. It would enable the use of large volumes of routinely acquired, unlabelled clinical imaging data (e.g. rare conditions and underrepresented populations) by relying solely on coarse, image- or volume-level labels. This paradigm shift would not only reduce the annotation burden but also offer a path toward more inclusive and scalable segmentation frameworks, less constrained by the biases and limitations of current labelled datasets (see Section 2.1.1).

One promising approach in this regard is ***anomaly detection***, a form of unsupervised learning designed to identify instances that markedly deviate from the typical characteristics of healthy anatomy (Baur et al., 2019, 2021; Fernando et al., 2021). Initial research utilised AEs trained exclusively on healthy data to learn a compact, lower-dimensional representation of healthy individuals. When abnormal or diseased data are passed through the AE, the model is unable to accurately reconstruct regions deviating from the learned data distribution. This leads to an increased reconstruction error in these regions, highlighting potential abnormalities and providing spatial information on the extent of the lesion (Atlason et al., 2019; Baur et al., 2019, 2021).

However, AEs are limited by their tendency to produce blurry reconstructions and their susceptibility to memorising the training data, which can lead to false positive predictions (Baur et al., 2019, 2021; X. Wu et al., 2021). To address these limitations, probabilistic generative models have been explored for anomaly detection, which estimate the underlying probability density of healthy anatomy. These generative models can be categorised by how they estimate the probability

density: ***likelihood-based models*** and ***implicit-generative models***. Likelihood-based models, such as variational autoencoders (VAEs) (Kingma & Welling, 2014), explicitly approximate the probability density function of healthy samples using maximum likelihood estimation. This enables the model to identify outliers as they fall outside the learned distribution. In contrast, implicit generative models such as GANs (Goodfellow et al., 2014) bypass explicit likelihood estimation by learning to generate realistic healthy samples through adversarial training (Schlegl et al., 2019; X. Wu et al., 2021).

Baur et al. (2019) combined adversarial training with a spatial VAE to overcome the memorisation limitations of AEs and improve reconstruction realism in low-resolution settings. Their model augments the standard reconstruction loss with an adversarial component, while regularising the encoder with a Kullback-Leibler (KL) divergence loss to encourage a well-structured latent space. However, as Zimmerer et al. (2019) note, the combined reconstruction and KL divergence losses merely approximate maximum likelihood and require manual hyperparameter tuning, which may limit generalisability. To improve robustness, they suggest directly incorporating the KL divergence into pixel-wise anomaly scoring.

To systematically compare AE- and GAN-based methods, Baur et al. (2021) evaluated several models, finding that GAN-based approaches such as f-AnoGAN and AnoVAEGAN produce more detailed reconstructions, particularly near anatomical boundaries. While AnoVAEGAN yielded sharper outputs, it suffered from anatomical inconsistencies and overfitting in datasets with small lesions. In contrast, f-AnoGAN offered better balance between image quality and structural coherence, outperforming other models in unsupervised anomaly detection. X. Wu et al. (2021) further improved performance by introducing anatomical priors through symmetry-based constraints, achieving segmentation results close to certain supervised baselines.

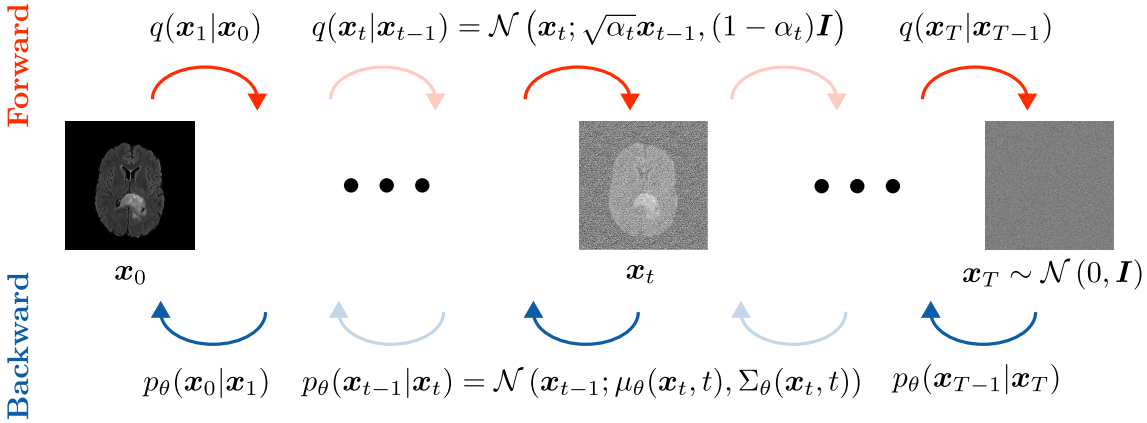


Figure 2.2: Schematic overview of the diffusion process. The forward process (top) progressively adds Gaussian noise to the original data point $\boldsymbol{x}_0$ over T steps, gradually removing information content. The backward process (bottom) is modelled by a NN parameterised by θ, which estimates the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ of the reverse transition distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. After training, novel samples are generated by traversing the learned backward process starting from isotropic Gaussian noise, $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$.

2.3 The importance of denoising diffusion models

Despite the demonstrated capabilities of VAEs and GANs in reflecting the underlying distribution of healthy data, both model categories encounter inherent limitations. Current likelihood-based models often produce detail-lacking blurry reconstructions due to their objective function, while GANs are challenging to train (Baur et al., 2021; Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021).

To address the limitations of current generative models, two distinct yet conceptually aligned approaches have been developed independently: ***diffusion models*** (Ho et al., 2020; Sohl-Dickstein et al., 2015; J. Song et al., 2021), and ***score-based generative models*** (Y. Song & Ermon, 2019). Both model families are built on the idea of perturbing data points from the input data distribution $q(\boldsymbol{x})$, reversing the perturbation by estimating the noise contribution or score function with a neural network (NN), and generating novel samples through iterative sampling. (Ho et al., 2020; Sohl-Dickstein et al., 2015; Y. Song & Ermon, 2019). The key difference between the two approaches lies in their objective functions: diffusion models are trained to estimate the conditional probability of the data at the next step given the

current step, while score-based models are trained to estimate the gradient of the log-probability density of the data distribution, commonly referred to as the score function (Ho et al., 2020; Y. Song & Ermon, 2019). Recent studies have shown that diffusion models offer advantages over score-based models in terms of theoretical simplicity, training stability, and sampling efficiency (S. Chen et al., 2022; Ho et al., 2020), and are therefore the focus of this literature review and research project.

2.3.1 Fundamental Components and Mechanisms

Since their introduction by Ho et al. (2020), DDPMs have become the dominant framework for diffusion-based generative modelling. Most subsequent work builds on this formulation or its refinements, which in turn evolved from the original diffusion process proposed by Sohl-Dickstein et al. (2015). DDPMs are generative models that approximate the underlying data distribution $q(\boldsymbol{x})$ by decomposing the learning task into two distinct phases: a ***forward process*** and a ***backward process*** (see Figure 2.2). Gaussian noise is incrementally added to a data sample $\boldsymbol{x}_i \sim q(\boldsymbol{x})$ in the forward process, gradually erasing its information content. The backward process aims to reverse this corruption during training and reconstruct the original (Dhariwal & Nichol, 2021; Ho et al., 2020; Kingma & Welling, 2014; Nichol & Dhariwal, 2021).

As this entire research project (Chapters 3 to 5) builds on the principles of DDPMs, it is essential to provide a detailed overview of their fundamental components and mechanisms. This section therefore outlines the core components of DDPMs, including detailed descriptions of the forward and backward processes, the generative sampling procedure, architectural design choices, diffusion parameters, and recent advances in training methodologies. Together, these elements form the foundation for understanding the capabilities and limitations of modern diffusion-based models, and their usage for anomaly detection to delineate brain tumours. For a more in-depth theoretical overview, the reader is referred to Bishop and Bishop (2024d), Dhariwal and Nichol (2021), Ho et al. (2020), and Nichol and Dhariwal (2021).

Forward Process

The forward process q is composed of a Markov Chain with T steps, which produces latent variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ given a data distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$. Each forward step $t \in \{1, \ldots, T\}$ adds Gaussian noise following a predefined noise schedule with variance $\beta_t \in (0, 1)$ as follows:

$$q(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T | \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) \tag{2.9}$$

$$q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I\right) \tag{2.10}$$

Given sufficient timesteps T, $\boldsymbol{x}_T$ will resemble an isotropic Gaussian distribution, i.e. zero mean and unit variance $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ (Ho et al., 2020; Nichol & Dhariwal, 2021). Utilising the reformulations $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and the reparametrisation trick

$$z = \mu + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0, \boldsymbol{I}), \quad z \sim \mathcal{N}(\mu, \sigma^2) \quad , \tag{2.11}$$

Equation (2.9) can be rewritten in closed form:

$$q\left(\boldsymbol{x}_t | \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}\right) \tag{2.12}$$

This allows the direct estimation of any $\boldsymbol{x}_t$ conditioned on $\boldsymbol{x}_0$ in the forward process (Ho et al., 2020; Kingma & Welling, 2014; Nichol & Dhariwal, 2021):

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \boldsymbol{I}) \tag{2.13}$$

Backward process: estimating the noise contribution

Since the forward process converges to an isotropic Gaussian, a novel data point $\hat{\boldsymbol{x}}_0 \sim q(\mathbf{x}_0)$ can be generated by first sampling from a standard Gaussian, $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$, and then reversing the forward process. However, $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ depends on the full, unknown data distribution $q(\mathbf{x})$ and a mixture of T Gaussian transitions, which is intractable to compute. To simplify estimation, Ho et al. (2020) model

the backward process as a conditional distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$. Conditioning on the noise-free data point $\boldsymbol{x}_0$ allows direct estimation of noise contributions at each timestep t. By observing the noisy sample $\boldsymbol{x}_t$ with information about its origin ($\boldsymbol{x}_0$), the model gains additional guidance for more accurate noise estimation. The conditional posterior distribution can be calculated using Bayes' theorem as follows (Bishop & Bishop, 2024d; Ho et al., 2020; Nichol & Dhariwal, 2021):

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{2.14}$$

The conditional posterior distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ takes the form of a Gaussian distribution and can be reformulated using the posterior mean $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$ and the posterior variance $\tilde{\beta}_t$ (Bishop & Bishop, 2024d; Ho et al., 2020):

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \boldsymbol{I}\right) \tag{2.15}$$

$$\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 \tag{2.16}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t \tag{2.17}$$

The reverse distribution is approximated using a NN, which estimates the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ of each step t of the backward process (Ho et al., 2020; Nichol & Dhariwal, 2021):

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \tag{2.18}$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)\right) \tag{2.19}$$

To avoid training T individual NNs, the timestep index t is included as an explicit input to the NN, enabling the timestep-dependent estimation of the mean and variance with a single NN. The timestep is hereby provided to the model using the sinusoidal position embedding of the transformer (Bishop & Bishop, 2024d; Ho et al., 2020; Nichol & Dhariwal, 2021; Vaswani et al., 2017).

As a maximum-likelihood model, the objective is to minimise the negative log-likelihood (NLL) of the data distribution, given by $-\log p_\theta(\boldsymbol{x}_0)$. Since $p_\theta(\boldsymbol{x}_0)$ is computationally intractable, the NLL is approximated using a surrogate objective, typically the variational lower bound (VLB). This approximation is feasible because the forward process can be viewed as a latent variable model, where the VLB serves as a tractable lower bound to the intractable marginal likelihood, analogous to the variational approach used in models like VAEs. The VLB is defined as (Bishop & Bishop, 2024d; Ho et al., 2020; Kingma & Welling, 2014):

$$\begin{aligned}\mathbb{E}\left[-\log p_\theta(\boldsymbol{x}_0)\right] &\leq \mathbb{E}_q\left[-\log p_\theta(\boldsymbol{x}_0) + D_{\mathrm{KL}}(q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0))\right] \\ &= \mathbb{E}_q\left[-\log\left(\frac{p_\theta(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\right)\right] \end{aligned} \tag{2.20}$$

The term D_{KL} denotes the KL divergence, which measures the discrepancy between the variational distribution $q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)$ and the model's approximate posterior $p_\theta(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)$. As a non-negative measure of information loss, it acts as a regularisation term, enforcing that the learned distribution approximates the true posterior. The VLB in Equation (2.20) can then be reformulated into the final NLL expression (Bishop & Bishop, 2024d; Ho et al., 2020; Sohl-Dickstein et al., 2015):

$$\begin{aligned}\mathbb{E}\left[-\log p_\theta(\boldsymbol{x}_0)\right] \leq\ \mathbb{E}_q\Bigg[&\underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))}_{L_T} \\ &+ \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))}_{L_{t-1}} \\ &\underbrace{-\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)}_{L_0}\Bigg]\end{aligned} \tag{2.21}$$

L_T can be omitted as it has no learnable parameters and represents a small constant if the forward process converges to an isotropic Gaussian $q(\boldsymbol{x}_T|\boldsymbol{x}_0) \approx \mathcal{N}(0, \boldsymbol{I})$. L_{t-1} defines the KL divergence between the approximated reverse distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ and the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$. Each term is Gaussian by construction of the forward process and can be derived in closed form. To simplify, Ho et al.

(2020) fixed the variance of the backward process to the forward noise variance, i.e. $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_0, t) \coloneqq \beta_t = \sigma_t^2$. Thus, the NN only predicts the mean, $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$, which streamlines training and lowers computational cost (Ho et al., 2020; Nichol & Dhariwal, 2021). The final form of L_{t-1} is given as

$$L_{t-1} = \mathbb{E}_q \left[\frac{1}{2\sigma_t^2} \left\| \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) \right\|^2 \right] + C \quad , \tag{2.22}$$

where C is a constant term that does not depend on the model parameters. Directly predicting $\tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0)$ with a NN is the most straightforward solution. However, Ho et al. (2020) further refined the training process utilising Equation (2.16) and Equation (2.13), leading to the reparametrisation of $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ as

$$\begin{aligned} \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) &= \tilde{\boldsymbol{\mu}}_t \left(\boldsymbol{x}_t, \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \right) \right) \\ &= \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \right) \quad , \end{aligned} \tag{2.23}$$

$\boldsymbol{\epsilon}_\theta$ estimates the forward noise $\boldsymbol{\epsilon}_t$ from the noisy sample $\boldsymbol{x}_t$, which simplifies Eq. (2.22) (Ho et al., 2020):

$$L_{t-1} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\| \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}_t, t \right) \right\|^2 \right] \tag{2.24}$$

The reformulation defined in Equation (2.24) not only represents a simplification of VLB but also draws the connection to the aforementioned generative score matching (Y. Song & Ermon, 2019) as it resembles Langevin dynamics (Ho et al., 2020). The final term of the VLB in Equation (2.21), L_0, is realised by an independent discrete decoder. The decoder calculates the mass of the distribution centred on the predicted mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_1, 1)$ with variance σ_1^2 in the interval of the real value of $\boldsymbol{x}_0 \pm 1/255$ for each pixel D (Ho et al., 2020):

$$p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) = \prod_{i=1}^{D} \int_{\delta_-(\boldsymbol{x}_0^i)}^{\delta_+(\boldsymbol{x}_0^i)} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_\theta^i(\boldsymbol{x}_1, 1), \sigma_1^2)\, \mathrm{d}x$$
$$\delta_+(x) = \begin{cases} \infty & \text{if } x = 1 \\ x + \frac{1}{255} & \text{if } x < 1 \end{cases} \qquad \delta_-(x) = \begin{cases} -\infty & \text{if } x = -1 \\ x - \frac{1}{255} & \text{if } x > -1 \end{cases} \tag{2.25}$$

If the NN predicts a similar pixel value to $\boldsymbol{x}_0$, the mass of the distribution for that individual pixel is high, resulting in a high contribution to the probability $p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1)$. The bounds in Equation (2.25) are derived from the preprocessing step, where each pixel value is rescaled from the range $[0, 255]$ to $[-1, 1]$, ensuring linearly scaled values that approximate an isotropic Gaussian distribution (Ho et al., 2020).

Generative process: Obtaining a new sample

The generative process of a new sample $\hat{\boldsymbol{x}}_0$ is realised by sampling from a standard Gaussian distribution $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ and transitioning the learned backward process of Equation (2.18). As each step of the backward process is defined by a Gaussian distribution (see Equation (2.19)), one can make use of the reparametrisation trick of Equation (2.11) and the parametrisation of $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ in Eq. (2.23) to generate $\boldsymbol{x}_{t-1}$:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \right) + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}) \tag{2.26}$$

During this process, Ho et al. (2020) noticed two useful simplifications to the framework resulting in better sampling quality and easier implementation: (1) re-weighting Equation (2.24) by removing the normalisation term, and (2) absorbing L_0 into the NN by setting the noise term of Equation (2.26) to 0 for $t = 0$. Both simplifications lead to the final objective of DDPM approximating the weighted VLB emphasising different properties of the standard diffusion process (Ho et al., 2020; Nichol & Dhariwal, 2021):

$$L_{\text{simple}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\| \boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \|^2 \right] \tag{2.27}$$

Equation (2.27) describes the mean-squared error (MSE) between the predicted noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ and the actual added noise during the forward process $\boldsymbol{\epsilon}_t$ (Ho et al., 2020; Nichol & Dhariwal, 2021).

Backbone architecture and diffusion parameters

The choice of the respective NN architecture to estimate the noise contribution, $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$, is primarily constrained by the requirement to retain dimensionality. Specifically, the predicted output of the NN must match the dimensionality of the input. As a result, the U-Net by Ronneberger et al. (2015) has been established as the gold-standard NN architecture used for DDPMs. Particularly in image applications, the CNN core of the U-Net achieves state-of-the-art feature extraction represented by sampling quality surpassing state-of-the-art GANs (Dhariwal & Nichol, 2021; Ho et al., 2020).

To estimate the time-dependent noise contribution, the current timestep t of the sample $\boldsymbol{x}_t$ is provided as input to the U-Net. This is done by using the sinusoidal position embedding from the transformer (Vaswani et al., 2017):

$$\begin{aligned} \mathrm{PE}_{(t,2i)} &= \sin\left(\frac{t}{10000^{\frac{2i}{d}}}\right) \\ \mathrm{PE}_{(t,2i+1)} &= \cos\left(\frac{t}{10000^{\frac{2i}{d}}}\right) \end{aligned} \tag{2.28}$$

d describes the feature dimensionality of the embedding, whereas the index i is designed to create unique positional encodings by alternating between sine and cosine. This allows the U-Net to distinguish between different timesteps by feeding the embedding to each individual block in encoder, bottleneck and decoder of the U-Net (Ho et al., 2020; Vaswani et al., 2017). The second alteration of the standard U-Net is the incorporation of the self-attention mechanism (Vaswani et al., 2017) (see Equation (2.8)), which enhances the model's feature extraction capabilities by increasing the receptive field at different resolutions (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021).

The diffusion process is characterised by the noise schedule β_t that defines the variance of the Gaussian noise added to the input data at each timestep. The choice of the noise schedule is crucial for the training process and the quality of the generated samples. Ho et al. (2020) proposed a linear noise schedule with $T = 1000$ ranging from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$, which has been determined empirically by Ho et al. (2020) and was selected for its ease of use and implementation.

Improvements to the training process

Since the introduction of diffusion models by Ho et al. (2020), researchers have actively pursued improvements to the training process due to the promising ability of DDPMs to model complex probability densities (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021).

The first major improvements to the DDPM framework were presented by Nichol and Dhariwal (2021) and focus on the VLB by proposing two key improvements: (1) Estimation of $\boldsymbol{\Sigma}_\theta$, and (2) adaptation of the noise schedule. The authors noted that fixing the variance $\boldsymbol{\Sigma}_\theta$ in L_{t-1} (see Equation (2.20)) compromises the log-likelihood approximation. This limitation stems from the dominance of the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}, t)$ over the variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ in determining the distribution with $T \to \infty$. To improve the log-likelihood, Nichol and Dhariwal (2021) include the estimation of the variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ into the training process. The variance is hereby estimated as an interpolation between forward process variance β_t and posterior variance $\tilde{\beta}_t$:

$$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \exp\left(v \log \beta_t + (1 - v) \log \tilde{\beta}_t\right) \tag{2.29}$$

The interpolation factor v is a learnable parameter of the NN and allows the model to adapt the variance estimation to the respective noise level of the forward process and the posterior distribution. The novel term of the VLB is then added to the simple loss function of Equation (2.27) as follows:

$$L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{VLB}} \tag{2.30}$$

The authors set $\lambda = 0.001$ to balance the loss terms and ensure variance estimation does not dominate noise estimation. The mean of L_{VLB} is subject to a stop-gradient operation to preserve the influence of L_{simple} on $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and provide a guidance term for $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ through LVLB (Nichol & Dhariwal, 2021).

The second improvement proposed by Nichol and Dhariwal (2021) is the adaptation of the noise schedule β_t to include more timesteps $T = 4000$ and to modify the noise schedule to a cosine annealing schedule. The authors noted that the linear noise schedule was suboptimal for low-resolution images as the tail of the schedule was too noisy without any meaningful image content. The revised noise schedule is defined as

$$\bar{\alpha}_t = \frac{f(t)}{f(0)}, \quad f(t) = \cos\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)^2 \quad . \tag{2.31}$$

The parameter $s = 0.008$ is estimated such that $\sqrt{\beta_0}$ is smaller than the pixel bin size. Utilising $\cos^2$ ensures that the destruction of information content is slower around the extrema of the diffusion process, i.e. $t = 0$ and $t = T$, whilst retaining the linear reduction of $\bar{\alpha}_t$ in the middle of the diffusion process (Nichol & Dhariwal, 2021).

One major drawback of DDPMs relative to GANs and VAEs is the substantially longer sampling time. The sample quality is directly correlated with the length of the diffusion process as larger T reduce the complexity of estimating $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ due to the simpler approximation of the Gaussian transition. As a result, the sampling process needs to sequentially traverse the entire Markov chain in reverse to generate a new sample, compared to the single-step sampling of GANs (Ho et al., 2020; Nichol & Dhariwal, 2021; Sohl-Dickstein et al., 2015; J. Song et al., 2021).

To address the slow sampling process, J. Song et al. (2021) generalised the forward Markov chain in DDPMs (see Equation (2.9)) to a non-Markovian diffusion. They observed that the loss in Equation (2.27) depends only on the marginals $q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ (see Equations (2.12) and (2.13)) rather than the full joint probability distribution $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$, enabling reformulation of the inference process:

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0}{\sqrt{1-\bar{\alpha}_t}}, \sigma_t^2 I\right) \tag{2.32}$$

parameterised by σ_t, which controls the stochasticity of each step in the forward process. The forward process can be derived from Equation (2.32) using Bayes formula (see Equation (2.14)) and the condition that $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is equal to Equation (2.12). The authors demonstrated that generalising the forward process introduces a novel variational inference objective, which resembles the objective of DDPMs for a given combination of $\sigma > 0$ and specific weights $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_T]$. These weights re-weight the loss term in Equation (2.27) for each timestep t. As a result, pre-trained DDPMs can be utilised for the generalised inference process, which approximate a family of non-Markovian forward processes in addition to the Markovian forward process of DDPMs (J. Song et al., 2021).

The generalised inference process is defined as a two step solution: given a noisy observation $\boldsymbol{x}_t$ obtained through the forward process (see Equation (2.12)), the backbone NN estimates the noise contribution $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ and utilises a re-formulation of Equation (2.12) to obtain a prediction of the corresponding $\boldsymbol{x}_0$. The latter is then leveraged in the reverse conditional distribution of Equation (2.32). This procedure redefines the sampling process of DDPMs generalised by σ_t using Equation (2.11) as follows (J. Song et al., 2021):

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{predicted } \boldsymbol{x}_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}_{\text{direction pointing to } \boldsymbol{x}_t} + \sigma_t \boldsymbol{\epsilon} \quad (2.33)$$

One particular special case of the generalised DDPM is the original DDPM with $\sigma_t = 0$ for all $t > 1$. The forward process becomes fixed given $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_0$, which results in a deterministic transition from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$ during the generative process. This model is referred to as denoising diffusion implicit model (DDIM). Equation (2.33) also entails the original generative formulation of the DDPM using $\sigma_t = \tilde{\beta}_t$.

DDIM reaches its main advantage in the accelerated sampling process. J. Song et al. (2021) observed that utilising a subset of diffusion steps $\{\boldsymbol{x}_{\tau_1}, \ldots, \boldsymbol{x}_{\tau_s}\}$ parameterised by an increasing subsequence $\tau \in [1, \ldots, T]$ of length S shortens the generative process. This is possible because the subsequence is constructed such that the

forward process marginals align with those of the full sequence, i.e. $q(\boldsymbol{x}_{\tau_i}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{\tau_i}}\boldsymbol{x}_0, (1-\alpha_{\tau_i})\boldsymbol{I})$. Depending on the respective choice of σ_t, the inference process can be markedly accelerated as only the shortened subsequence needs to be traversed. With this modification, DDIMs achieve accelerations of up to 50x with limited loss in detail compared to the DDPM due to the deterministic and more predictable sampling process mapping directly from noisy inputs to samples (J. Song et al., 2021).

2.3.2 Latent diffusion: training and inference with reduced dimensionality

Rombach et al. (2022) took a different approach to reduce the computational complexity associated with the iterative Markov chain process. The authors leveraged pre-trained AEs to encode each sample $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)$ into a latent space representation $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_0)$ using the encoder $\mathcal{E}$ of the AE. The latent representation is then used as input to the regular training process of the DDPM. This includes obtaining a noisy observation $\boldsymbol{z}_t$ via Equation (2.13), where $\boldsymbol{x}_0$ is replaced by the encoded latent $\boldsymbol{z}_0$. During the backward process, the noise contribution $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t)$ is estimated in the latent space, optimising the VLB using Equation (2.27). To obtain a novel sample $\hat{\boldsymbol{x}}_0$, the generative process is traversed in the latent space, and the latent space sample $\hat{\boldsymbol{z}}_0$ is decoded back to the original sample space using the decoder $\mathcal{D}$ of the AE, such that $\hat{\boldsymbol{x}}_0 = \mathcal{D}(\hat{\boldsymbol{z}}_0)$ (Rombach et al., 2022). As the diffusion model operates exclusively within the latent space of the AE, it is termed a latent diffusion model (LDM) (Rombach et al., 2022).

The computational complexity is negatively correlated to the compression factor of the AE with compression factors of 4 to 16 demonstrating exceptional generative results in the tested configuration of Rombach et al. (2022). The authors provided an additional theoretical formulation for the observed detail preservation in scenarios with reduced complexity. The majority of the information content of the image corresponds to imperceptible details. The utilisation of the AE eliminates these details during the encoding process, and allows the diffusion model to concentrate

on the semantic and perceptual composition during training. As a result, generated images retain the essential features required for perceptually high-quality samples with a fraction of the computational cost (Rombach et al., 2022).

Rombach et al. (2022), building on the work of Esser et al. (2021), investigate two methods for regularising the latent space in the first-stage model. The first approach, autoencoder with Kullback-Leibler regularisation (KLAE), imposes a KL divergence penalty to encourage a Gaussian latent distribution, similar to a VAE. In contrast, the autoencoder with vector quantisation regularisation (VQAE) aims to discretise the latent space using a learnable embedding. Both first-stage models are accompanied by a discriminator for adversarial training. Rombach et al. (2022) demonstrated that the KLAE and VQAE models with mild compression factors of 4 to 16 yield great results in terms of perceptual quality and sample diversity using $T = 1000$, with minor differences between both results for different tasks.

Khader et al. (2023) extended the LDM framework to the generation of 3D medical data by following the pathway of the VQAE approach. In addition, the authors employed a perceptual loss to improve reconstruction quality, leveraging pre-trained perceptual models such as LPIPS (see Section 2.1.2). To adapt the framework for 3D data, Khader et al. (2023) replaced each 2D layer with its 3D counterpart, maintaining the design principles of the original implementation. Since LPIPS is trained on 2D ground-truth perceptual datasets, Khader et al. (2023) applied a slice-based perceptual loss to approximate the advantages of LPIPS in 3D.

Relation to thesis

This thesis employs LDMs for their efficiency and flexibility in modelling high-resolution medical images, enabling faster generation with fewer timesteps and reduced computational demands.

2.3.3 Conditional DDPMs

Conditional density estimation is a fundamental concept in generative DL, where the goal is to model the distribution of $\boldsymbol{x}$ given auxiliary task-related information

$\boldsymbol{y}$. By incorporating this conditioning signal, the model restricts its focus to the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, effectively reducing the complexity of the learning problem. This targeted approach enables the model to learn context-dependent distributions that reflect prior knowledge encoded in $\boldsymbol{y}$ (e.g. class labels, textual descriptions, or reference images) thereby producing outputs that are better aligned with specific tasks or constraints (Dhariwal & Nichol, 2021; Rombach et al., 2022).

As described in Section 2.3.1, DDPMs are already conditioned on the original sample $\boldsymbol{x}_0$ and the timestep t during training. This section focuses on the conditioning of DDPMs on additional information, such as class labels or other images, which are of importance for subsequent applications, such as conditional image generation, or super-resolution (SR) in Chapter 4.

Class-conditional diffusion models

Dhariwal and Nichol (2021) proposed a widely adopted method for class-conditioning in DDPMs, introducing a learnable embedding $\boldsymbol{y} = \boldsymbol{L}$ for each class label $\boldsymbol{L}$. This embedding is injected into each residual block through "adaptive group normalisation", allowing the model to modulate intermediate activations based on the target class. The DDPM is thus trained to estimate the conditional distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{L})$ at each diffusion timestep t, enabling the synthesis of samples that exhibit class-specific characteristics.

Generalised conditioning mechanism

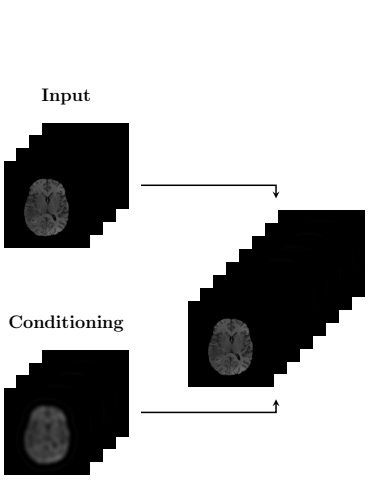
Rombach et al. (2022) generalised the conditioning strategy beyond single-dimensional class labels to various auxiliary information sources, including text embeddings, images and semantic information. Instead of utilising the adaptive group normalisation to inject the conditioning signal into the U-Net, Rombach et al. (2022) devised a modality-specific encoder τ_ϕ to compute an intermediate representation $\tau_\phi(\boldsymbol{y})$ of the conditional input $\boldsymbol{y}$. The encoder τ_ϕ is jointly optimised with the U-Net using Equation (2.27), where $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is extended with the additional conditioning signal $\tau_\phi(\boldsymbol{y})$:

$$L_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(\boldsymbol{x}),\boldsymbol{y},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \tau_\phi(\boldsymbol{y}))\|_2^2\right] \tag{2.34}$$

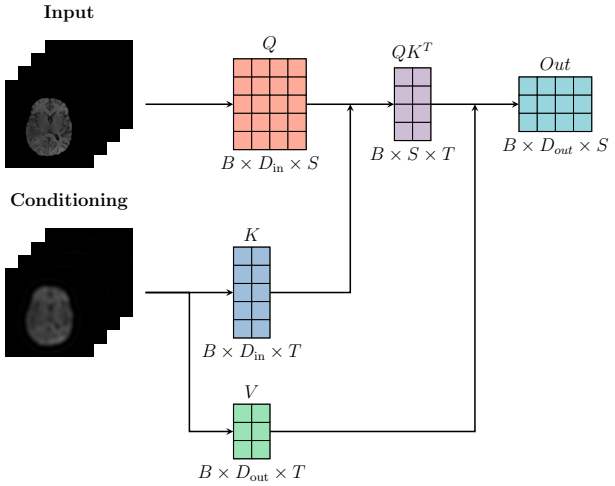
This framework offers a straightforward yet versatile means of incorporating additional conditioning signals to guide the generative process of DDPMs (Rombach et al., 2022).

The encoder τ_ϕ can be realised in various ways, depending on the type of conditioning signal and the desired output. The easiest method is the concatenation of the conditioning signal (see Fig. 2.3a) with the input $\boldsymbol{x}_t$ at the encoder stage, which can be used for inputs of the same dimensionality (Rombach et al., 2022). This entails imaging data and semantic information, such as segmentation masks or anatomical structures. However, this approach may not be suitable for all types of conditioning signals, especially in the case of LDMs operating on a compressed input space. Unless the encoder of the compression stage $\mathcal{E}$ is trained on the same distribution of data for conditioning and input, ***cross-attention*** (see Fig. 2.3b) is required to align the conditioning signal with the input data (C.-F. R. Chen et al., 2021; Rombach et al., 2022). Cross-attention introduces additional learnable layers to enable multi-scale feature fusion between the conditioning signal and the input data, allowing for more complex interactions and relationships to be captured (C.-F. R. Chen et al., 2021; Rombach et al., 2022).

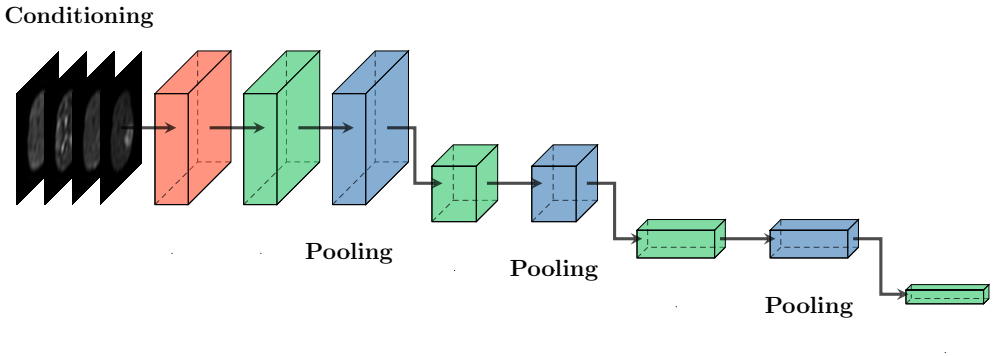
Since the introduction of the conditioning concept, there has been an increasing interest in enhancing the feature density of the conditioning signal. Specifically, Park et al. (2019) observed that earlier strategies may cause the guidance signal to vanish through subsequent normalisation layers, resulting in a loss of information. To address this, they introduced SPatially-Adaptive (DE)normalization (SPADE) to preserve spatial conditioning signals by modulating normalisation parameters through learned, spatially varying affine transformations. These feature extractors can be included in conventional architectures and allow deep feature injection (Park et al., 2019). This approach has sparked the conceptualisation of distinct encoders mimicking the structure of the U-Net (see Fig. 2.3c), which can be used to increase the feature



(a) Channel-wise concatenation



(b) Cross-attention



(c) Encoder-based conditioning

Figure 2.3: Visualisation of conditioning strategies using a low-resolution version of the input as the conditioning signal. (a) shows channel-wise concatenation, which requires the input and conditioning signal to share identical spatial dimensions. (b) illustrates cross-attention conditioning, which allows differing spatial resolutions (S, T) by operating on flattened spatial embeddings, with original resolution restored after conditioning. (c) depicts encoder-based conditioning with deep feature extraction: features from each pooling block are passed through a secondary extractor (not shown) and injected into the SPADE blocks of the backbone U-Net. Features are reused across layers with matching spatial resolution.

extraction capabilities of conditioning signals. Architectures utilising mirrored U-Net encoders are employed to exploit previously obtained priors with lower computational overhead compared to traditional cross-attention conditioning (Patil et al., 2025; J. Wang et al., 2024; L. Zhang, Rao, & Agrawala, 2023). A prominent example is the

ControlNet architecture, which injects spatial priors into a frozen base model using additive zero-convolutions (L. Zhang, Rao, & Agrawala, 2023). In contrast, J. Wang et al. (2024) leverage the computational efficiency of SPADE-based modulation to inject features into pre-trained LDMs. This approach facilitates high-resolution image synthesis with improved feature density and structural preservation compared to the additive mechanism of ControlNet or the expensive cross-attention mechanism (see Section 2.1.3).

In summary, DDPMs employ three principal conditioning strategies, which directly influence the choice of τ_ϕ:

1. ***channel-wise concatenation***, offering a straightforward and computationally efficient approach for conditioning inputs of identical dimensionality (see Figure 2.3a);
2. ***cross-attention***, enabling learnable conditioning by handling inputs with differing spatial resolutions (see Figure 2.3b); and
3. ***encoder-based conditioning***, which leverages deep feature extraction and pre-trained models to efficiently integrate spatial conditioning (see Figure 2.3c).

Relation to thesis

Conditional sampling underpins the anomaly detection framework introduced in Section 2.3.6, which forms the basis of **this thesis**. The conditioning strategies beyond injection of class information are applied in Chapter 4 to generate targeted small-lesion samples and improve spatial resolution.

2.3.4 Medical image generation with DDPMs

Synthetic data generation is a crucial task in medical imaging, enabling both the augmentation of training datasets to address class imbalance and the creation of standardised samples with predefined characteristics that are otherwise challenging to obtain from real patient data (Fernandez et al., 2024; Meng et al., 2024; W. Peng et al., 2023; Sizikova et al., 2024; H. Wu et al., 2024; Yi et al., 2019). As DDPMs are fundamentally designed as generative models, they have been increasingly applied

to the generation of synthetic medical images, with $\boldsymbol{y}$ being usually constrained to class labels or image-level guidance to control the appearance of the synthetic sample (Dorjsembe et al., 2024; Fernandez et al., 2022, 2024; H. Wu et al., 2024). This section provides an overview of the application of DDPMs in medical image generation, focusing on the conditioning strategies employed to obtain samples with specific characteristics, particularly lesion sizes and locations.

Synthetic medical image generation is a well-established field, with approaches targeting modality restoration (Kalantar et al., 2023; Meng et al., 2024; Müller-Franzes et al., 2023), domain translation (Graf et al., 2023; X. Li et al., 2023; Patil et al., 2025), and conditioning on auxiliary inputs such as text (Kim et al., 2024; H. Wu et al., 2024; Z. Zhang et al., 2024). These methods rely on conditioning mechanisms within DDPMs (see Section 2.3.3).

Fernandez et al. (2022, 2024) used a two-stage pipeline where a conditional LDM generates segmentation masks, which are then fed into a SPADE-based VAE-GAN to synthesise brain images. H. Wu et al. (2024) conditioned abdominal MRI synthesis on organ masks and text, using a dedicated mask encoder aligned with the input latent space to improve spatial consistency. Similarly, Dorjsembe et al. (2024) used binary lesion masks via channel-wise concatenation to guide a DDPM in image space, achieving a 5% Dice score improvement when training segmentation models on generated samples. Konz et al. (2024) extended this binary mask conditioning approach to breast and abdominal imaging, showing negligible segmentation performance differences between real and synthetic data, supporting the anatomical fidelity of their samples. Their model, like that of Bhattacharya et al. (2025), incorporated fine-grained anatomical labels to further improve structural consistency in the generated outputs.

Relation to thesis

These studies show that LDMs can generate synthetic medical images with targeted lesion characteristics using simple conditioning inputs like binary masks. **This thesis** applies this principle in Section 4.3 to synthesise a dataset of small brain tumours, enabling controlled evaluation across varying lesion characteristics.

2.3.5 DDPMs for super-resolution

Relation to thesis

Detecting small lesions remains one of the key challenges in medical imaging and is directly tied to early diagnosis and improved clinical outcomes (see Section 2.1.6). However, such lesions are often difficult to identify due to limited spatial resolution and imaging artefacts. SR offers a promising strategy to mitigate this issue by enhancing the visibility of subtle pathological features. **In the context of this thesis**, it is hypothesised that applying SR to synthetic datasets of small lesions, as mentioned in the previous section and demonstrated in Section 4.3, may improve detection performance by refining spatial detail.

The following section introduces the theoretical foundations of SR with DDPMs and discusses the construction of paired high- and low-resolution datasets, which are essential for training these models effectively (Moser et al., 2025; Shao et al., 2023; Yang et al., 2023; H. Zhou et al., 2022).

Theoretical foundations of super-resolution

SR encompasses a range of techniques aimed at enhancing the spatial resolution of an image, enabling the reconstruction of high-resolution (HR) images, $\boldsymbol{I}_{\mathrm{HR}}$, from low-resolution (LR) inputs, $\boldsymbol{I}_{\mathrm{LR}}$. This process typically involves approximating a degradation function $\mathcal{D}$ that captures the relationship between the HR and LR images (Moser et al., 2025; Shao et al., 2023; Yang et al., 2023; H. Zhou et al., 2022):

$$\boldsymbol{I}_{\mathrm{LR}} = \mathcal{D}(\boldsymbol{I}_{\mathrm{HR}}) \tag{2.35}$$

$\mathcal{D}$ denotes the loss of overall image quality inherent to the imaging modality, typically linked to blurring, noise, or compression artefacts. The goal of learning-based SR is to approximate the inverse of the degradation function, $\hat{\boldsymbol{I}}_{\mathrm{HR}} = \mathcal{D}^{-1}(\boldsymbol{I}_{\mathrm{LR}}, \theta)$, using learnable parameters θ. This mapping estimates the HR image $\hat{\boldsymbol{I}}_{\mathrm{HR}}$ from the LR input $\boldsymbol{I}_{\mathrm{LR}}$ and is expressed as the following optimisation objective (Moser et al., 2025; Shao et al., 2023; Yang et al., 2023; H. Zhou et al., 2022):

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{I}_{\mathrm{LR}}, \boldsymbol{I}_{\mathrm{HR}}} \left[\mathcal{L}(\boldsymbol{I}_{\mathrm{HR}}, \mathcal{D}^{-1}(\boldsymbol{I}_{\mathrm{LR}}, \theta)) \right] \quad . \tag{2.36}$$

Generating paired HR-LR datasets

Obtaining a dataset with paired HR-LR images is essential for training SR models and their objective defined in Equation (2.36). Since capturing the exact same scene at multiple resolutions is challenging even in natural imaging, early approaches instead relied on obtaining HR images and applied a predefined synthetic degradation function $\mathcal{D}_s$ to generate LR counterparts. This is particularly relevant for medical applications, where obtaining a LR image often requires altering the imaging protocol, increasing acquisition time and cost, and patient discomfort (Shin et al., 2024; Yang et al., 2023; H. Zhang et al., 2023; H. Zhou et al., 2022). In addition, clinical MRIs demonstrate large variability in terms of scanner settings, acquisition protocols, and patient-specific factors, further complicating the acquisition process (Thust et al., 2018; Villanueva-Meyer et al., 2017; J. Wang et al., 2023). Initial approaches successfully generated SR data by inverting fixed perturbations defined in $\mathcal{D}_s$, but were inherently limited by the narrow scope of these predefined degradations. As a result, applying $\mathcal{D}_s^{-1}$ to data exhibiting unseen or mismatched degradations led to poor reconstruction quality, highlighting the critical need to accurately model the complex and diverse degradation processes present in real-world clinical data (H. Chen et al., 2021; Dai et al., 2019; J. Wang et al., 2024; Z. Zhang et al., 2019).

Subsequent work shifted focus towards degradation processes that more accurately reflect real-world conditions (Fritsche et al., 2019; L. Wang et al., 2021; Wei et al., 2021), or towards the construction of paired HR-LR datasets for supervised training (X. Wang et al., 2021; K. Zhang et al., 2021). Among these, the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) degradation pipeline proposed by X. Wang et al. (2021) has emerged as a state-of-the-art in natural image SR research. This pipeline enhances model generalisation by simulating a diverse range of degradation types commonly observed in practical imaging scenarios (Saharia et al., 2023; J. Wang et al., 2024). It applies a two-stage process involving Gaussian blurring, downsampling, the addition of various noise types, and compression. Due to the partially stochastic and non-deterministic nature of this degradation procedure, the corresponding SR task is termed ***blind SR***.

DDPMs for blind super-resolution

Demonstrated by Rombach et al. (2022) and described in detail in Section 2.3.3, DDPMs are capable of learning the distribution of HR images conditioned on a LR image. By injecting structural information from the LR image during training using the encoder $\tau_\phi(\boldsymbol{I}_{\text{LR}})$, the model implicitly learns the inverse degradation function $\mathcal{D}_s^{-1}$ as part of the standard DDPM training framework (see Section 2.3) (Moser et al., 2025; Rombach et al., 2022):

$$p_\theta(\boldsymbol{x}_{0:T}) = p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{I}_{\text{LR}}) \tag{2.37}$$

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{I}_{\text{LR}}) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{I}_{\text{LR}}), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t, \boldsymbol{I}_{\text{LR}})\right) \tag{2.38}$$

Saharia et al. (2023) introduced SR3, a diffusion-based framework for SR, which conditions the generation process on an upsampled LR image concatenated with the noisy HR image. Despite exploring more complex conditioning strategies, the simple upsampling-based approach proved sufficient. SR3 also employs a cascaded design, training separate models for each resolution scale to incrementally refine

image quality and avoid the challenges of direct large-scale upsampling.

Rather than training from scratch, J. Wang et al. (2024) employ a pre-trained LDM as a generative prior for SR, integrating multi-scale features from a dedicated LR encoder via SPADE (see Section 2.3.3). The encoder mirrors the U-Net structure but is trained separately, while the pre-trained generative model remains fixed. This design lowers training complexity and surpasses baseline methods.

Relation to thesis

The approach by J. Wang et al. (2024) is compelling for SR, combining pre-trained LDMs with a task-specific encoder to reduce training complexity. Its modular architecture and strong performance mark it as a promising state-of-the-art method. **This thesis** explores its applicability to brain MRI in Section 4.4, investigating whether this architecture can enhance the visibility of small tumour lesions and support high-sensitivity clinical imaging.

Super-resolution in MRI

Given the substantial differences in image composition between natural and medical images, specialised models are required to enhance resolution and detail in medical imaging applications. These differences encompass intricate anatomical structures, higher intrinsic noise levels, varying noise characteristics, and variable contrasts, which stem from the underlying imaging modalities (Aja-Fernández & Vegas-Sánchez-Ferrero, 2016; Shao et al., 2023; Shin et al., 2024). Consequently, SR approaches developed for natural images often fail to generalise effectively to medical imaging, as they do not account for modality-specific artefacts and the unique statistical properties of medical data (H. Chung et al., 2023; Shin et al., 2024). As MRI is the gold-standard for brain imaging (see Section 1.1.1), it is essential to develop SR methods that are tailored to the specific characteristics of MRI data, which are outlined in the following.

H. Chung et al. (2023) propose a score-based reverse diffusion framework for MRI denoising and SR, which refines corrupted images through iterative sampling. This

approach mitigates over-smoothing and preserves high-frequency detail, offering improved efficiency due to the contracting nature of the score-based mapping. The denoiser is also adapted for SR tasks.

Mao et al. (2023) target multi-contrast SR using a channel attention-based fusion mechanism. Distinct encoders extract modality-specific features, which are adaptively integrated via attention to enhance image fidelity. Combined with curriculum learning and a tailored loss, the method outperforms conventional SR models.

J. Wang et al. (2024) and J. Wang et al. (2023) introduce an optimisation-based approach for anisotropic SR, leveraging a pre-trained LDM. Starting from a random latent $\boldsymbol{z}_T$, the model iteratively refines this variable by comparing its downsampled reconstruction to the original LR input, optimising $\boldsymbol{z}_T$ with respect to the known degradation function $\mathcal{D}$.

H. Zhang et al. (2023) propose a self-supervised framework for anisotropic MRI SR, repurposing existing MRI volumes to generate synthetic training pairs. Reformulating the 3D SR task as a 2D problem, the model learns from high-resolution planar images and aggregates axial outputs at inference to reconstruct isotropic volumes.

Adaptations to the degradation pipeline The distinct characteristics of MRI require custom definitions of $\mathcal{D}_s$ to ensure effective SR performance (Lepcha et al., 2023; Shao et al., 2023; H. Zhou et al., 2022). To this end, Han et al. (2023) propose a probabilistic degradation model that learns the degradation function rather than relying on static assumptions. Their model is trained on both natural and medical images, transferring knowledge from natural image degradations to improve generalisation in the medical domain. Alternatively, Shao et al. (2023) adapt the ESRGAN degradation pipeline by restricting blur and noise types while randomising degradation order to better capture the stochastic nature of real-world artefacts (see Algorithm 1). In contrast, H. Zhou et al. (2022) estimate the degradation function in an unsupervised manner using unpaired HR-LR samples from different distributions. HR images are first downsampled with a predefined function, after which a Cycle-GAN refines the degraded images to match the LR domain. This

Algorithm 1 Adapted ESRGAN degradation pipeline for MRI by Shao et al. (2023).

Require: High-resolution MRI tensor $x \in \mathbb{R}^{H \times W}$
Require: Number of degradation levels N
Ensure: Degraded image tensor $\tilde{x}$ at original resolution
1: $\tilde{x} \leftarrow x$
2: **for** $i = 1$ to N **do**
3: $\quad \tilde{x} \leftarrow \text{Blur}(\tilde{x}, k_i)$ ▷ Blur Kernel $k_i \in \{\text{iso}, \text{aniso}\} \times \{\text{regular}, \text{generalised}, \text{plateau}\}$
4: $\quad \tilde{x} \leftarrow \text{Resize}(\tilde{x}, \text{scale} = s_i)$ ▷ Resize scale $s_i \in [0.3, 1.5]$
5: $\quad \tilde{x} \leftarrow \text{Noise}(\tilde{x}, \eta_i)$ ▷ Noise parameters $\eta_i \in \{\text{Gaussian, Poisson}\}$
6: **end for**
7: $\tilde{x} \leftarrow \text{Resize}(\tilde{x}, \text{target size} = H \times W)$ ▷ Resample to original resolution
8: **return** $\tilde{x}$

allows the model to approximate degradation components in a learnable manner, conceptually similar to B. Huang et al. (2021).

Relation to thesis

While the outlined degradation pipelines partially account for MRI-specific considerations, they remain approximations of real-world acquisition conditions. Most existing approaches overlook critical sources of variability in clinical imaging, such as complex noise distributions, motion-induced artefacts, and low-frequency intensity inhomogeneities introduced by bias fields. These omissions may limit the realism and generalisability of synthetic LR data. To address this problem, **this thesis** proposes an adapted degradation pipeline of Algorithm 1 in Section 4.4.1, which is specifically tailored to MRI and informed by the dominant artefacts encountered in practice.

The following sections detail the key factors influencing the design of the final degradation pipeline, including the incorporation of realistic noise models, simulated motion artefacts, and spatial bias fields. These are essential components for constructing more accurate and clinically meaningful degradation processes.

Noise distributions in MRI MRI is inherently susceptible to various types of noise, influenced by acquisition parameters and scanner hardware. In many practical

settings, noise is approximated as real-valued, additive Gaussian noise applied to the magnitude image:

$$\boldsymbol{I}_{\mathrm{LR}} = \boldsymbol{I}_{\mathrm{HR}} + \mathcal{N}(0, \sigma^2) \quad , \tag{2.39}$$

where $\mathcal{N}(0, \sigma^2)$ denotes zero-mean Gaussian noise with variance σ^2. While convenient, this model assumes post-reconstruction corruption and neglects the complex-valued nature of MRI signals (Aja-Fernández & Vegas-Sánchez-Ferrero, 2016; Gudbjartsson & Patz, 1995).

In reality, noise originates in the complex domain, where real and imaginary components follow $S_r, S_i \sim \mathcal{N}(0, \sigma^2)$. After magnitude reconstruction, the signal $M = \sqrt{(S_r + A)^2 + S_i^2}$ deviates from a Gaussian distribution, especially at low signal-to-noise ratio. A more accurate representation is the Rician distribution:

$$p(M \mid \boldsymbol{I}_{\mathrm{HR}}, \sigma) = \frac{M}{\sigma^2} \exp\left(-\frac{M^2 + \boldsymbol{I}_{\mathrm{HR}}^2}{2\sigma^2}\right) B^{(0)}\left(\frac{\boldsymbol{I}_{\mathrm{HR}} M}{\sigma^2}\right) \quad , \tag{2.40}$$

where $B^{(0)}$ is the zeroth-order modified Bessel function of the first kind, and $\boldsymbol{I}_{\mathrm{LR}} \sim p(M \mid \boldsymbol{I}_{\mathrm{HR}}, \sigma)$ denotes a sample from this distribution.

For modern multi-coil systems, image reconstruction uses a root-sum-of-squares approach across L independent channels, leading to a non-central Chi distribution with $2L$ degrees of freedom:

$$p(M \mid \boldsymbol{I}_{\mathrm{HR}}, \sigma, L) = \frac{\boldsymbol{I}_{\mathrm{HR}}^{1-L}}{\sigma^2} M^L \exp\left(-\frac{M^2 + \boldsymbol{I}_{\mathrm{HR}}^2}{2\sigma^2}\right) B^{(L-1)}\left(\frac{\boldsymbol{I}_{\mathrm{HR}} M}{\sigma^2}\right) \quad , \tag{2.41}$$

where $B^{(L-1)}(\cdot)$ is the modified Bessel function of the first kind of order $L - 1$ (Aja-Fernández & Vegas-Sánchez-Ferrero, 2016). Differences in the resulting noise characteristics are illustrated in Figs. B.5b and B.5c and Section B.2.

Bias fields MRI is commonly affected by low-frequency intensity non-uniformities known as bias fields. These arise from magnetic field inhomogeneities, spatially varying coil sensitivities, and patient-specific anatomical positioning, and manifest

as smooth intensity variations that confound subsequent image analysis (Sled et al., 1998; Sudre et al., 2017; Van Leemput et al., 1999). The observed image $\boldsymbol{I}_{\mathrm{LR}}$ can be expressed as a multiplicative combination of the true underlying image $\boldsymbol{I}_{\mathrm{HR}}$ and a spatially smooth bias field B_F:

$$\boldsymbol{I}_{\mathrm{LR}} = \boldsymbol{I}_{\mathrm{HR}} \times B_F \quad . \tag{2.42}$$

In practice, B_F is often modelled as a linear combination of low-degree polynomial basis functions (Sudre et al., 2017; Van Leemput et al., 1999). The bias-field degradation is shown in Fig. B.5e, with clearly visible intensity inhomogeneities.

Motion artefacts in MRI Patient motion during acquisition introduces artefacts in MRI, commonly manifesting as blurring or ghosting. These artefacts result from spatial inconsistencies in k-space caused by motion during the relatively long acquisition period and can substantially degrade diagnostic image quality (Dale et al., 2015; Shaw et al., 2019).

Rigid body motion can be modelled in the Fourier domain as phase shifts in k-space, with translational motion corresponding to deterministic displacements via the Fourier shift theorem (Shaw et al., 2019). This allows simulation of motion artefacts through controlled k-space manipulations. In contrast, physiological motion (e.g. respiration, cardiac pulsation, or cerebrospinal fluid flow) produces ghosting artefacts along the phase-encoding direction, which may appear as periodic signal duplications or diffuse noise depending on the regularity of motion (Axel et al., 1986; Storey et al., 2002; Wood & Henkelman, 1985). These effects are illustrated in Figs. B.5f and B.5g.

As no motion-related artefacts were observed in the BraTS data, they were not included in the final degradation pipeline. However, the functionality is implemented and can be adapted for datasets where such artefacts are present.

2.3.6 Anomaly detection with denoising diffusion models

The synthetic generation capabilities of DDPMs also allow the detection of anomalies in medical imaging. The majority of these approaches are based on the assumption that the generative model can learn the underlying data distribution of healthy training data, which allows it to generate a healthy counterfactual of the input. Anomalies are then detected by measuring the difference between the generated sample and the input (Behrendt et al., 2024; Fontanella et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022).

Healthy counterfactual generation

Drawing inspiration from reconstruction-based VAEs (Baur et al., 2021), one possible approach is to train the DDPM solely on healthy data. This constrains the model to generate a healthy counterpart of the input, facilitating direct comparison for the identification of anomalous regions. However, there are two limitations to this approach: (1) absence of publicly available, state-of-the-art datasets comprising exclusively healthy individuals, and (2) inability to accurately detect anomalous regions with low-intensity deformations.

As outlined in Section 2.1.1, the state-of-the-art dataset for brain tumour segmentation encompasses exclusively pathological individuals. Other collections containing healthy individuals are limited by the available MRI sequences, which are crucial to obtain the complete image of the lesion (Carrete et al., 2022; IXI Dataset, 2004–2006; Villanueva-Meyer et al., 2017). Secondly, Meissen et al. (2022) demonstrated that models trained exclusively on healthy data are only able to detect anomalies if they are of high-intensity, which exhibits similar or worse performance compared to simple thresholding. Sanchez et al. (2022) argue deformations of healthy tissue as a result of a macroscopic lesion are challenging to detect without the presence of anomalous examples in the training data. As a result, the authors propose to train the DDPM on a mixture of healthy and anomalous data to improve the detection of anomalies. Their findings demonstrated the inferior segmentation performance

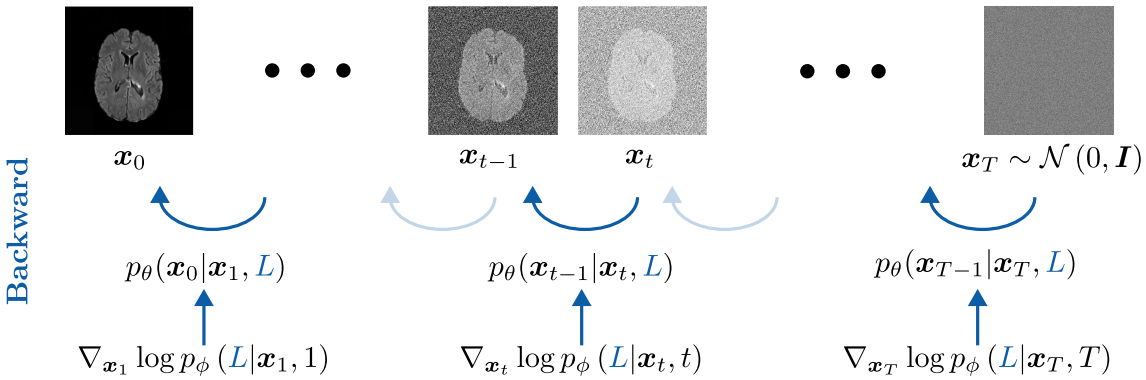


Figure 2.4: Conditional generative process of a DDPM with class information conditioned on $\boldsymbol{L}$ = "healthy". The reverse process begins from Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ and iteratively reconstructs a sample by injecting class information at each timestep t. Conditioning is applied via adaptive group normalisation and through gradient guidance using $\nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t)$.

of DDPMs trained exclusively on healthy data compared to models trained in a weakly-supervised manner with a mixture of healthy and anomalous data (Sanchez et al., 2022).

In addition to injecting class information into the U-Net (see Section 2.3.3), Dhariwal and Nichol (2021) proposed to utilise an auxiliary classifier $p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t)$ to increase the guidance of the diffusion model during sampling (see Figure 2.4), and therefore the generation of healthy counterfactuals for anomaly detection. The classifier is hereby trained on noisy samples $\boldsymbol{x}_t$ to predict the associated label $\boldsymbol{L}$. The respective gradient $\nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t)$ can be used to guide the generation towards the desired class label $\boldsymbol{L}$. The impact of the gradient on the sampling is controlled by a scaling factor C. Dhariwal and Nichol (2021) reformulated specific components of the sampling process. Stochastic conditional sampling can be performed using $p_{\theta,\phi}(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}, \boldsymbol{L}) = Z \cdot p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1}) p_\phi(\boldsymbol{L}|\boldsymbol{x}_t)$ following Bayes' theorem with Z describing a normalisation constant. Exact sampling from this distribution is intractable, which requires the approximation of the distribution using a distorted Gaussian distribution. The latter is derived through a Taylor expansion around the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$, which results in the altered mean shown in Equation (2.43). The stochastic re-formulation does not apply to DDIMs as they are deterministic models. Instead, DDIMs resort

to score-based conditioning, leveraging the connection between diffusion models and score-based networks to enable conditional sampling (Dhariwal & Nichol, 2021):

$$\text{DDPM:} \quad \hat{\boldsymbol{\mu}}_\theta(\boldsymbol{x}_t, t, \boldsymbol{L}) = \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t) + s \cdot \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) \cdot \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t) \tag{2.43}$$

$$\text{DDIM:} \quad \hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, \boldsymbol{L}) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \sqrt{1 - \bar{\alpha}_t} \cdot \nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t) \tag{2.44}$$

The conditional sampling proposed by Dhariwal and Nichol (2021) requires training of an auxiliary classifier to obtain $\nabla_{\boldsymbol{x}_t} \log p_\phi(\boldsymbol{L}|\boldsymbol{x}_t, t)$, which increases the computational complexity of the approach. In addition, these classifiers are task-specific as they have to be trained on the noisy data precluding the usage of conventional pre-trained classifiers. As a result Ho and Salimans (2021) introduced the concept of ***classifier-free guidance*** by embedding the classifier into the DDPM. This approach leverages the label embedding and the encoding of label information $\boldsymbol{L}$ into the backbone U-Net (see Section 2.3.3) to simultaneously train two diffusion models: a conditional model $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{L})$ and an unconditional model $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{L} = \emptyset)$. The unconditional model is obtained by randomly masking the label information $\boldsymbol{L}$ and the subsequent class embedding during training. The conditional model steers the process towards the target class without requiring an explicit classifier, resulting in improved control over the generation process and reduced computational overhead. This approach effectively trains two individual models using a singular architecture and a unified training process. Both models can then be used to modify the noise prediction $\boldsymbol{\epsilon}(\boldsymbol{x}_t, t)$ during generation similar to Equations (2.43) and (2.44) with classifier guidance scale C (Ho & Salimans, 2021):

$$\hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_t, t) = (1 + C) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{L}) + C \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{2.45}$$

Relation to thesis

Classifier-free guidance is central **to this thesis**, as it enables the generation of healthy counterfactuals without the need for an auxiliary classifier. This approach is central to diffusion-based anomaly detection and is employed throughout Chapters 3 to 5 to generate healthy counterparts of diseased data.

Latent space encoding

As outlined in Section 2.3.1, the generative process begins with a noisy sample $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ and produces an output that reflects general characteristics of the target distribution $\boldsymbol{L}$ through conditional sampling. However, this output captures only broad features of healthy anatomy and does not preserve the subject-specific anatomical structure of the input image $\boldsymbol{x}_0$. As a result, it fails to produce a synthetic healthy counterfactual tailored to the original input.

J. Song et al. (2021) noticed that DDIM with $\sigma_t = 0$ can be utilised to encode the observation $\boldsymbol{x}_0$ into the latent space. Particularly, the authors noticed the resemblance of the reverse DDIM step defined in Equation (2.33) to the Euler method for solving ordinary differential equations

$$\sqrt{\frac{1}{\bar{\alpha}_{t-1}}}\boldsymbol{x}_{t-1} = \sqrt{\frac{1}{\bar{\alpha}_t}}\boldsymbol{x}_t + \left(\sqrt{\frac{1-\bar{\alpha}_{t-1}}{\bar{\alpha}_{t-1}}} - \sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}}\right) \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$$
$$\boldsymbol{x}_{t+\Delta t} \approx \boldsymbol{x}_t + \Delta t \cdot f(\boldsymbol{x}_t, t) \leftarrow \text{Euler method} \tag{2.46}$$

with Δt describing the step size of the Euler method and $\sigma_t = 0$. If T is large enough, the individual step sizes Δt become very small, which results in an almost perfect approximation by the Euler method to solve the ordinary differential equation. In this scenario, the Euler method describes the true continuous trajectory of the system, which allows the approximation of forward step being equal to the reverse step, i.e. $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \approx \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_{t-1}, t)$. The learned backwards process can therefore be used to encode the input into the latent space of the diffusion process, allowing healthy

counterfactual generation (Dhariwal & Nichol, 2021; Fontanella et al., 2024; Sanchez et al., 2022; J. Song et al., 2021; Wolleb et al., 2022):

$$\boldsymbol{x}_{t+1} = \sqrt{\bar{\alpha}_{t+1}} \left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t+1}} \cdot \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{2.47}$$

Wallace et al. (2023) noticed that the approximation of Eq. (2.47) is inherently fragile, as reconstructions of unaltered image regions often fail to replicate the original input accurately. Furthermore, stronger conditional alterations exacerbate these reconstruction errors, degrading image quality and limiting flexibility, particularly for shorter diffusion sequences (Hertz et al., 2023; Wallace et al., 2023). To mitigate the limitations of the DDIM encoding mechanism, Wallace et al. (2023) propose a reformulation based on affine coupling layers (ACLs) (Dinh et al., 2015, 2017).

ACLs encode an input $\boldsymbol{z}$ into a latent space while avoiding the linearisation assumption of the diffusion process, ensuring exact invertibility. This property is critical for achieving reversible transformations, enabling precise reconstruction and robust editing. In the affine coupling mechanism, the input $\boldsymbol{z}$ is split along the channel dimension into two equal parts, $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$. Using two neural networks, ψ and Ψ, the transformed output $\boldsymbol{z}'_a$ is computed as follows (Wallace et al., 2023):

$$\boldsymbol{z}'_a = \Psi(\boldsymbol{z}_b)\boldsymbol{z}_a + \psi(\boldsymbol{z}_b) \tag{2.48}$$

Importantly, the process is invertible: the original $\boldsymbol{z}$ can be reconstructed from the overall output $\boldsymbol{z}' = [\boldsymbol{z}'_a, \boldsymbol{z}_b]$ of the ACL by reversing the transformation of Eq. (2.48):

$$\boldsymbol{z}_a = \frac{\boldsymbol{z}'_a - \psi(\boldsymbol{z}_b)}{\Psi(\boldsymbol{z}_b)} \tag{2.49}$$

This invertibility is essential for the diffusion process, as it ensures the original input can be recovered without error. Specifically, ACLs track two quantities ($\boldsymbol{z}_a$ and $\boldsymbol{z}_b$) that invert each other, enabling precise encoding and decoding.

Building on this concept, Wallace et al. (2023) reformulate the sampling process in DDIM, assuming $\sigma_t = 0$ (see Eq. (2.33)):

$$\boldsymbol{z}_{t-1} = a_t \boldsymbol{z}_t + b_t \epsilon(\boldsymbol{z}_t, t) \quad . \tag{2.50}$$

Here, a_t and b_t are time-dependent coefficients chosen for simplicity and given by

$$a_t = \frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}} \tag{2.51}$$

$$b_t = -\sqrt{\frac{\bar{\alpha}_{t-1}(1-\bar{\alpha}_t)}{\bar{\alpha}_t}} + \sqrt{1-\bar{\alpha}_{t-1}} \quad . \tag{2.52}$$

Eq. (2.50) becomes fully affine in $\boldsymbol{z}_t$ and $\boldsymbol{\epsilon}$ by redefining $\boldsymbol{\epsilon}(\boldsymbol{z}_t, t)$ to be independent of $\boldsymbol{z}_t$ through the substitution $\boldsymbol{y} = \boldsymbol{z}_t$. The substitution resembles Eq. (2.48) and allows the definition of simple update rules for the reverse process (Wallace et al., 2023):

$$\begin{aligned} \boldsymbol{z}_{t-1} &= a_t \boldsymbol{z}_t + b_t \boldsymbol{\epsilon}(\boldsymbol{y}_t, t) \\ \boldsymbol{y}_{t-1} &= a_t \boldsymbol{y}_t + b_t \boldsymbol{\epsilon}(\boldsymbol{z}_{t-1}, t) \end{aligned} \tag{2.53}$$

These update rules allow efficient forward and backward traversal of the latent space, preserving invertibility. Notably, while the method utilises the linearisation $\boldsymbol{\epsilon}(\boldsymbol{x}_t, t) \approx \boldsymbol{\epsilon}(\boldsymbol{x}_{t-1}, t)$, it does not rely on it for invertibility. This ensures exact recovery of the original input and allows for precise edits demonstrated in natural imaging data (Wallace et al., 2023), reflected in the naming of the method as exact diffusion inversion via coupled transformations (EDICT).

Stabilisation with mixing layers Wallace et al. (2023) observed that $\boldsymbol{z}_t$ and $\boldsymbol{y}_t$ diverge considerably using small numbers of encoding steps $1 < N << T$. This divergence is caused by the breakdown of the linearisation assumption, which becomes more pronounced with fewer encoding steps. To address this issue, the authors introduce mixing layers that compute a weighted average of both noise vectors, with the mixing weight $\omega \in [0, 1]$:

$$
\begin{aligned}
\boldsymbol{z}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}/\bar{\alpha}_t}\boldsymbol{z}_t + \sqrt{1-\bar{\alpha}_{t-1}} * \boldsymbol{\epsilon}(\boldsymbol{y}_t, t) \\
&\quad - \sqrt{\bar{\alpha}_{t-1} * (1-\bar{\alpha}_t)/\bar{\alpha}_t} * \boldsymbol{\epsilon}(\boldsymbol{y}_t, t) \\
\boldsymbol{y}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}/\bar{\alpha}_t}\boldsymbol{y}_t + \sqrt{1-\bar{\alpha}_{t-1}} * \boldsymbol{\epsilon}(\boldsymbol{z}_{t-1}, t) \\
&\quad - \sqrt{\bar{\alpha}_{t-1} * (1-\bar{\alpha}_t)/\bar{\alpha}_t} * \boldsymbol{\epsilon}(\boldsymbol{z}_{t-1}, t) \quad . \qquad (2.54)
\end{aligned}
$$

Relation to thesis

The EDICT encoding mechanism, originally proposed for natural images, enables stronger conditional edits within fewer diffusion steps. Its impact in medical imaging remains unexplored and is investigated as part of **this thesis** in Chapter 3, where its potential for efficient and controllable synthesis is evaluated in the context of brain tumour segmentation.

Brain tumour segmentation with denoising diffusion models

Recent advancements in diffusion models for medical imaging have explored various strategies to enhance anomaly detection and segmentation performance. These studies build upon foundational diffusion frameworks and introduce novel adaptations tailored to specific challenges in medical data.

Wolleb et al. (2022) applied the DDPM framework with an auxiliary classifier and U-Net to guide sampling towards healthy anatomy, achieving a Dice score of 0.71 via Otsu thresholding on central 2D slices. Sanchez et al. (2022) improved this approach by training on both healthy and diseased slices, adopting classifier-free guidance (Ho & Salimans, 2021), and introducing dynamic normalisation to avoid saturation. Their method reached a Dice of 0.76 on BraTS 2021 undersampled slices.

Pinaya, Tudosiu, et al. (2022) adopted a 2D LDM (Rombach et al., 2022) trained on healthy data, detecting anomalies via deviations in noise prediction across timesteps and achieving a Dice of 0.40, indicating the inferiority of unsupervised anomaly detection. Wyatt et al. (2022) proposed simplex noise to replace Gaussian corruption, improving anomaly detection for large lesions and shorter denoising chains. Behrendt

et al. (2024) introduced a patch-based DDPM with implicit conditioning and direct sampling, reaching a Dice of 0.49 using simplex noise.

Finally, Graham et al. (2023) extended these ideas to 3D, demonstrating that high-fidelity latent representations are critical for out-of-distribution detection on synthetic anomalies and CT data, outperforming latent transformer baselines.

2.4 Identified research gaps of weakly-supervised brain tumour segmentation

Supervised DL methods are widely regarded as the state-of-the-art for medical image segmentation, often achieving remarkable performance metrics. However, these methods face major limitations stemming from their dependence on large-scale, high-quality annotated datasets. The acquisition of such datasets is constrained by several factors. Ethical considerations and institutional policies at hospitals frequently hinder the collection and sharing of medical imaging data at scale. Additionally, the annotation process is resource-intensive and ultimately costly, as it requires domain-specific expertise to manually label each individual across potentially numerous MRI acquisitions with complementary sequences. The annotation process itself introduces additional challenges. Annotator bias, whether due to variability in expert interpretations or systematic differences in labelling practices, may inadvertently impose skewed representations of disease characteristics. Furthermore, supervised methods are inherently disease-specific, limiting their generalisability to other pathological conditions or cases not represented in the training data.

To address these challenges, there has been a growing interest in how to reduce the degree of supervision required. One promising direction relies on anomaly detection by modelling healthy anatomical distributions. In this context, DDPMs have recently emerged as a powerful alternative to conventional generative models, demonstrating strong potential in anomaly detection tasks. The following sections identify key research gaps in the current state-of-the-art of weakly-supervised brain tumour segmentation with DDPMs, which this thesis aims to address.

2.4.1 Overcoming computational and data limitations in weakly-supervised 3D brain tumour segmentation

Gap 1

The transition from 2D to 3D medical image analysis for weakly-supervised brain tumour segmentation remains underexplored due to increased computational demands and the lack of public datasets with healthy individuals.

State-of-the-art models employing DDPM for weakly-supervised brain tumour segmentation may exhibit constrained representational capacity due to their reliance on lower-dimensional architectures. Notably, the predominant paradigm favours 2D implementations of DDPM for tumour detection (Behrendt et al., 2024; Fontanella et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022). This stands in contrast to the increasing clinical adoption of high-resolution 3D MRI sequences (Carrete et al., 2022; Villanueva-Meyer et al., 2017), which offer superior spatial fidelity. The preference for 2D methodologies over their 3D counterparts can primarily be attributed to the following factors:

1. increased computational complexity associated with 3D operations, which substantially extends both training and inference times,
2. easier extraction of healthy anatomy from diseased individuals, and
3. the capacity to leverage and adapt pre-existing DDPM frameworks for integration into weakly-supervised segmentation pipelines.

Studies suggest that 3D approaches hold substantial promise for advancing medical image analysis, particularly in tasks like brain tumour segmentation, as they capture the inter-slice relationships of anatomical structures and lesion distribution, providing a more detailed spatial context (Avesta et al., 2023; Kazerooni et al., 2024; Singh et al., 2020; X. Zhou et al., 2018). However, 3D image analysis incur higher computational costs due to higher dimensionality of the data and, more importantly, the increased number of parameters of the learnable feature extractors of the backbone CNN. This results in higher memory consumption and longer training times as part of the

more challenging parameter optimisation. Limited computational resources may also prevent the training of a model with sufficient capacity to learn the complex patterns in the data. This is particularly crucial for the conventional attention mechanism embedded in the backbone U-Net of the DDPM framework, as its complexity grows quadratically with the number of input features. As shown in Section 2.1.3, the input features are directly influenced by the dimensionality of the data (Rabe & Staats, 2022).

Studies suggest that 3D approaches hold substantial promise for advancing medical image analysis, particularly in brain tumour segmentation, as they capture inter-slice relationships of anatomical structures and lesion distribution, providing richer spatial context (Avesta et al., 2023; Kazerooni et al., 2024; Singh et al., 2020; X. Zhou et al., 2018). However, 3D analysis incurs higher computational costs due to the data's dimensionality and the larger number of parameters in backbone CNNs. This leads to greater memory consumption and longer training times from more demanding parameter optimisation. Limited computational resources may also prevent training models with sufficient capacity to capture complex patterns. The issue is especially critical for the attention mechanism embedded in the U-Net backbone of the DDPM. Its complexity scales quadratically with the number of input features, which are directly influenced by data dimensionality (see Section 2.1.3) (Rabe & Staats, 2022).

The second factor influencing the decision to use 2D models arises from the requirements of weakly-supervised brain tumour segmentation and the limited availability of public datasets. The BraTS dataset, the most widely used dataset for brain tumour segmentation (see Section 2.1.1), consists exclusively of pathological individuals, lacking healthy subjects for comparison. This absence of healthy data poses a major challenge for weakly-supervised approaches, which rely on examples of healthy anatomy to identify and localise anomalous regions effectively. To address this issue, recent work has proposed to extract 2D slices with absent ground-truth annotations from diseased volumes to serve as substitutes for healthy anatomy (Behrendt et al., 2024; Fontanella et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022). Naturally, transitioning from 2D to 3D introduces the challenge

of extracting healthy anatomy from diseased volumes. A potential solution involves incorporating private datasets of healthy subjects; however, this approach is severely constrained by ethical considerations, as outlined in Section 2.1.1, and is incompatible with the BraTS challenge, which prohibits the use of private data collections (Bakas et al., 2017, 2018; Menze et al., 2015). An alternative approach is to adapt the extraction of healthy anatomy specifically for 3D, bypassing the need for private data while maintaining accuracy.

The final constraint limiting the transition to 3D is the prevalence of publicly available DDPM frameworks, which are primarily designed for 2D applications. This bias reflects the abundance of natural image datasets and the rapid progress of 2D generative modelling (Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021; Rombach et al., 2022; J. Song et al., 2021). Although existing 2D frameworks can be adapted with domain-specific features, their limited modularity restricts substantial modifications beyond their original task and dimensionality. Such rigidity constrains their applicability to medical imaging, where domain-specific challenges demand tailored solutions for optimal performance (Cardoso et al., 2022; Cheplygina, 2019; Isensee et al., 2021; Llambias et al., 2024).

In addition to the challenges posed by dimensionality choices, the encoding mechanism used to preserve anatomical features during healthy counterfactual generation plays a pivotal role in the success of weakly-supervised brain tumour segmentation. This mechanism is essential for maintaining the integrity of healthy regions while enabling modifications to pathological areas during the reconstruction of a healthy counterpart. State-of-the-art methods rely on deterministic reformulations of the diffusion process in DDIMs, which leverage a linearisation assumption to encode anatomical details into the latent noise vector (Sanchez et al., 2022; J. Song et al., 2021; Wolleb et al., 2022) (see Section 2.3.6). However, recent work (Wallace et al., 2023) has highlighted fundamental flaws in the underlying linearisation assumption of the DDIM encoding mechanism, demonstrating that it often leads to suboptimal encoding, reduced fidelity in image modifications, and, ultimately, ineffective conditional modifications in practice. To address these shortcomings, Wallace et al.

(2023) introduced EDICT to offer a novel framework that circumvents the limitations inherent to DDIMs. While EDICT has shown considerable improvements in conditional sampling for natural images, its application to medical imaging remains largely unexplored. Given the potential advantages of this method in enhancing the quality and stability of generative models, this project aims to investigate the applicability of the EDICT encoding mechanism to improve weakly-supervised brain tumour segmentation. Notably, EDICT offers pronounced modifications with as few as 50 steps (Wallace et al., 2023) and has the ability to further reduce inference times in the computationally expensive 3D DDPM process. This enables more efficient and effective segmentation, addressing both performance and computational concerns in medical imaging tasks.

To address Gap 1, I synthesised the following research question, which is the focus of **Chapter 3**:

Research Question 1

How can weakly-supervised 3D anomaly detection be made computationally efficient while mitigating reliance on publicly available datasets of healthy individuals?

2.4.2 The impact of super-resolution on weakly-supervised segmentation of small brain tumours

Gap 2

The role of SR in enhancing weakly-supervised DDPM-based anomaly detection remains underexplored, particularly with respect to its effect on sensitivity to small or subtle brain tumours.

Early and accurate detection of brain tumours is particularly critical in paediatric populations, where delays in diagnosis can lead to irreversible developmental and neurological impairments (Alemany et al., 2021; Fry et al., 2014; Goldman et al., 2017; Sadighi et al., 2018; Yamada et al., 2020). The developing brain is highly

vulnerable to treatment-induced neurotoxicity, especially during periods of rapid growth (de Ruiter et al., 2013; Palmer et al., 2001; Ris et al., 2001), and prolonged diagnostic timelines have been associated with long-term deficits in cognitive, sensory, and motor function, with over 60% of childhood cancer survivors experiencing life-long disabilities (Armstrong, 2010; Lassaletta et al., 2015; Wilne et al., 2010). While timely imaging depends on broader clinical infrastructure (Sadighi et al., 2018; Walker et al., 2016), definitive diagnosis ultimately relies on the radiological identification of small, early-stage lesions (Sabeghi et al., 2024; Wilne et al., 2010). These lesions, which often represent the earliest detectable manifestations of tumour development, pose a major diagnostic challenge due to their subtle appearance and strong dependence on imaging protocols (M. Chen et al., 2022; L. Zhang, Wen, et al., 2023). Despite the clinical relevance of small lesion detection, current DL models for brain tumour segmentation are rarely assessed in terms of size-specific performance. In particular, the BraTS dataset (Bakas et al., 2017) primarily features large, macroscopic lesions, rendering it unsuitable for benchmarking model efficacy on small tumours.

Existing DL methods for small lesion detection have predominantly targeted pathologies that are inherently small: lung nodules (Abraham & Khan, 2019; Lin et al., 2020; Y. Zhang et al., 2024), liver tumours (Savelli et al., 2020), or stroke lesions (An et al., 2023; Tao et al., 2019; Wong et al., 2022). These approaches often employ supervised learning and benefit from task-specific innovations such as focal loss and scale-aware architectures (Luo et al., 2024). As a result, these approaches are subject to the same limitations as outlined in Section 2.1.8. Weakly-supervised methods for small lesion detection remain underexplored, likely because the task represents a worst-case scenario: detecting subtle deviations from the healthy distribution without any label guidance poses a major challenge. Even in fully supervised settings, performance degrades with decreasing lesion size (Erdur et al., 2024; L. Li et al., 2021; B. Xu et al., 2018), highlighting the intrinsic difficulty of this task.

One promising avenue for improving small lesion detection involves to increase the spatial resolution and fidelity of the input data. Medical imaging data is often acquired at limited resolution and is prone to modality-specific artefacts, which can obscure subtle pathological changes. These limitations pose challenges not only for clinical interpretation but also for automated detection methods, particularly when attempting to identify early-stage or small-scale abnormalities (Lepcha et al., 2023; Luo et al., 2024; Shin et al., 2024). Enhancing the resolution has the potential to reveal finer anatomical structures that are critical for early diagnosis. In this context, DDPM-based SR has emerged as a promising method for improving spatial fidelity (H. Chung et al., 2023; J. Wang et al., 2023) (see Section 2.3.5), though its effectiveness for small brain tumour detection remains unexplored - especially within weakly-supervised anomaly detection frameworks built on DDPMs. Existing pipelines have yet to establish a clear relationship between increased resolution and improved detection performance. Bridging this gap is essential for advancing clinically meaningful segmentation of early-stage tumours.

To address Gap 2, I synthesised the following research question, which is the focus of **Chapter 4**:

Research Question 2

What is the effect of DDPM-based SR on the sensitivity and segmentation performance of weakly-supervised anomaly detection, particularly for small brain tumours?

2.4.3 Exploring distributional overlap for generalisable brain tumour segmentation

Gap 3

The generalisability of weakly-supervised DDPMs to paediatric brain tumours remains untested, despite their hypothesised robustness to distributional shifts and suitability for data-scarce clinical settings.

Finally, the application of DL-based segmentation to paediatric brain tumours remains severely limited due to the lack of large-scale, annotated datasets. This scarcity, driven by ethical, institutional, and logistical constraints, precludes the training of supervised models from scratch, which typically require extensive labelled data to perform reliably (see Section 2.1.8). As a result, most existing segmentation frameworks have been developed and validated exclusively on adult populations. In this context, weakly-supervised anomaly detection using DDPMs offers a promising alternative. These models are hypothesised to generalise well under distributional shifts, yet their applicability to paediatric neuro-oncology has not been previously explored. This chapter presents the first systematic investigation of whether models pre-trained on adult data can be effectively applied to paediatric cases and evaluates the role of fine-tuning under limited-data conditions. In doing so, it advances the use of diffusion-based anomaly detection for paediatric brain tumour segmentation and assesses its robustness to cross-population shifts.

To address Gap 3, I synthesised the following research question, which is the focus of **Chapter 5**:

Research Question 3

To what extent can DDPMs trained on adult brain tumour data generalise to the paediatric domain, and how robust are the learned representations to shifts in population and disease distribution?

Chapter 3

Latent diffusion model for 3D brain tumour segmentation

Parts of this chapter have previously been published in the Proceedings of the 23rd International Conference of Artificial Intelligence in Medicine. See Publication 1 for more details.

This chapter introduces a novel patch-based latent diffusion model (LDM) for 3D weakly-supervised brain tumour segmentation, addressing the methodological gaps outlined in Section 2.4.1. By operating in the latent space, the proposed LDM markedly lowers the computational cost of the diffusion process, enabling the efficient modelling and analysis of volumetric data. In addition, this compact representation permits shorter diffusion sequences as a result of the feature-rich representation. To further accelerate inference and facilitate stronger conditional edits in fewer diffusion steps, a novel encoding strategy based on the exact diffusion inversion via coupled transformations (EDICT) framework is employed. The entire chapter is focused on demonstrating the proof-of-concept of a novel modular LDM framework, while emphasising efficiency in resource-constrained segmentation tasks.

The chapter is structured as follows: Section 3.1 motivates the development of a 3D-LDM and outlines the objectives of the chapter. Section 3.2 describes the conceptualisation of the model framework, including the data processing pipeline and the model architecture. This section utilises 2D natural images to verify that the

core diffusion and denoising logic are sound and capable of reproducing established benchmarks before the system is scaled. Section 3.3 subsequently extends the framework to 3D volumetric medical data and weakly-supervised brain tumour segmentation. The experimental results are presented in Section 3.4, including various ablation experiments. The chapter concludes with a summary of outstanding limitations in Section 3.5 and the components of the overarching contribution in Section 3.6:

Contribution 1

Developed a patch-based LDM for efficient 3D weakly-supervised brain tumour segmentation via anomaly detection, preserving volumetric context and integrating a robust encoding strategy to improve anatomical fidelity.

3.1 Introduction

As highlighted in Section 2.4.1, the majority of approaches utilising denoising diffusion probabilistic models (DDPMs) for weakly-supervised brain tumour segmentation rely on 2D implementations, despite the increasing volumetric nature of tomographic medical imaging data (Carrete et al., 2022; Villanueva-Meyer et al., 2017). This chapter aims to address this limitation by introducing a novel 3D-LDM framework for weakly-supervised brain tumour segmentation in magnetic resonance imaging (MRI). The proposed approach investigates a patch-based strategy in combination with the resource-efficient LDM to mitigate the computational requirements of 3D operations and enable the benefits of volumetric analysis. Furthermore, the patch-based approach allows to extract healthy subregions from otherwise diseased individuals as tumours are typically localised to specific areas of the brain. This enables the utilisation of pseudo-healthy training data derived from the same cohort of patients and alleviates the need for a separate healthy control group. To enhance the segmentation performance, this chapter investigates the EDICT encoding mechanism (see Section 2.3.6) as an alternative to the conventional denoising diffusion implicit model (DDIM) mechanism (see Section 2.3.6). EDICT is designed to improve the

capacity of the diffusion model to perform image edits through conditional sampling by modifying the encoding mechanism of the reverse sampling process. The approach has demonstrated strong performance in natural images, but its application to medical imaging remains largely unexplored. As a result, this chapter evaluates its benefits in the context of medical imaging and weakly-supervised brain tumour segmentation. Lastly, the overarching focus of this chapter is to demonstrate the proof-of-concept of a novel modular framework building on the theoretical foundations of DDPM that allows seamless expansion for subsequent chapters. As a result, the chapter is governed by the following two objectives:

Objective 1a: Modular model framework

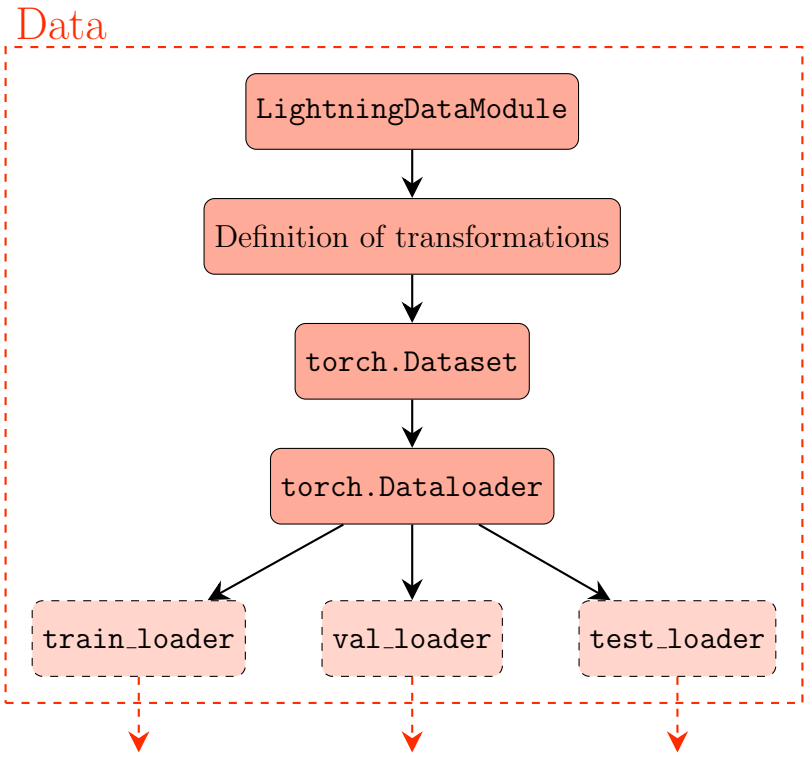
Develop a modular framework for DDPM-based brain tumour segmentation, enabling straightforward adaptation to varying dimensionalities, tasks, and backbone architectures through reusable components.

Objective 1b: LDM for 3D segmentation

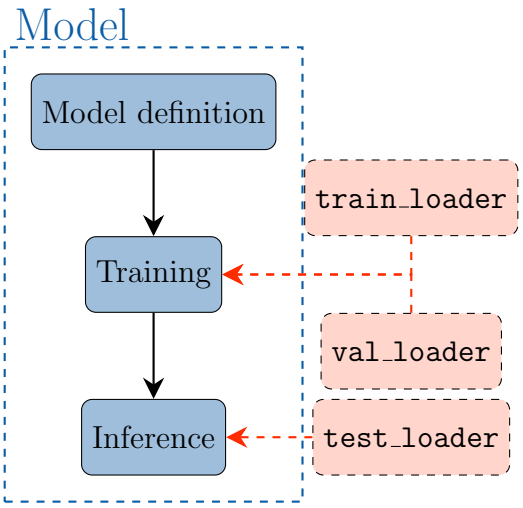
Develop a weakly-supervised LDM for efficient 3D brain tumour anomaly detection, incorporating patch-based sampling and robust encoding to reduce computational cost and mitigate reliance on external healthy data.

3.2 Conceptualisation of model framework

The initial focus is to establish a foundational model framework for weakly-supervised brain tumour segmentation in MRI scans. This framework comprises two key components: the ***data processing pipeline*** and the ***model architecture***. The framework is designed to be modular and adaptable to the specific requirements of medical imaging data. The primary goal is to develop a functional pipeline that can be systematically evaluated and validated, ensuring the successful integration of all components. This approach provides a solid foundation for the incremental expansion of the framework, allowing for seamless integration of advanced features



(a) Data processing pipeline



(b) Model architecture and training pipeline

Figure 3.1: Schematic overview of the data processing and model training pipelines. (a) illustrates dataset composition, preprocessing transformations, and integration with `torch` and `pytorch_lightning`. (b) outlines the training and evaluation procedures, following model definition and its associated components.

and customised adaptations tailored to diverse applications within the medical domain. An exact implementation of the framework and detailed descriptions of the components are provided in Section A.1.

3.2.1 Data processing pipeline

The ***data processing pipeline*** is designed to address the specific challenges of medical imaging data, ensuring efficient preparation for model training. Its primary goals are to handle multi-sequence MRI volumes, apply domain-specific preprocessing, and augment limited datasets through targeted transformations. The pipeline standardises data formats, integrates optional patient-level information, and extracts relevant spatial patches used during training. A key component is the patch sampling mechanism, which selects balanced samples of healthy and pathological regions to support the weakly-supervised learning framework (see Section 3.3.2). This ensures that the model is exposed to diverse anatomical variations while maintaining focus on lesion localisation. Augmentations are applied selectively to enhance data diversity without compromising anatomical integrity. By leveraging modular, task-specific transformations, the pipeline remains adaptable to different datasets and problem formulations, providing a robust foundation for efficient training. The backend for these operations is built on the `torchio` library (Pérez-García et al., 2021), which offers a comprehensive suite of tools for medical image processing, including loading, preprocessing, and augmentation functionalities tailored to 3D medical imaging data.

3.2.2 Backbone model architecture and diffusion backend

The ***model architecture*** follows a modular U-Net design, optimised for learning the reverse diffusion process necessary for healthy counterfactual generation (see Section 2.3). It integrates convolution and attention layers to extract multi-scale features essential for accurate reconstruction. Conditioning is incorporated via timestep and class embeddings, enabling the model to guide the denoising trajectory with temporal

and contextual information required for conditional sampling (see Section 2.3.6). The diffusion backend complements this by defining the forward corruption process and managing the reverse generative steps, abstracting the underlying diffusion mechanics from the model's learning objective. Together, these components form a cohesive framework capable of learning probabilistic mappings from noisy inputs to clean reconstructions, which is critical for effective anomaly detection in medical imaging.

3.2.3 Validation of model framework

Building on the framework described in the previous section, this experiment aims to systematically evaluate each component using a well-established backbone architecture and dataset. The goal is to replicate the original findings of Ho et al. (2020) within the proposed modular framework to validate its capacity for probability density estimation in natural images. This initial focus on 2D benchmarks provides a computationally efficient environment to verify that the core diffusion and denoising logic are sound. By confirming that the implementation can reproduce established results on natural images, the framework's reliability is established before transitioning to the 3D medical domain, where the increased dimensionality and data complexity make such iterative verification substantially less feasible. A successful replication serves as a functional verification of the implementation and supports the framework's reproducibility, laying the groundwork for its subsequent extension to the medical domain.

Dataset

The CIFAR-10 dataset (Krizhevsky & Hinton, 2009), consisting of 60,000 natural images across 10 distinct classes, is utilised for this preliminary proof of concept. It was chosen as it facilitates direct comparison to the results of Ho et al. (2020), Nichol and Dhariwal (2021), and Rombach et al. (2022), while being compact and resource-efficient, with image sizes of 32×32. This enables rapid identification of errors in the processing pipeline and facilitates quick iterations on the model architecture.

The pre-defined training split of 50,000 images is utilised for training and validation, with the images normalised to $[-1, 1]$ following standard procedure. During training, the images are augmented with random horizontal flips. The preprocessing pipeline follows the same steps as the original implementations but integrates the new data processing pipeline fundamentally designed for medical imaging data.

Model architecture and evaluation

The model architecture is implemented based on the configuration established by Ho et al. (2020). This includes a modular U-Net backbone architecture with 4 layers, 2 residual blocks per layer, and attention layers at a spatial resolution of 16×16. The diffusion process employs $T = 1000$ timesteps, with the variance β_t increasing linearly within the interval $\beta_t \in [0.0004, 0.02]$. The model is evaluated using the Fréchet inception distance (FID) score (see Section 2.1.2) by comparing 50,000 artificially generated samples using the fully trained network with the real samples of the training dataset. To streamline the application, `torchmetrics` (Detlefsen et al., 2022) is used, which is a `PyTorch`-based library for computing various evaluation metrics, including FID. This library is designed to be modular and extensible, allowing for easy integration into existing `pytorch-lightning` pipelines.

Experimental results

After several iterations, the final model achieves an FID score of 5.14, which is marginally higher than the value of 3.17 reported by Ho et al. (2020). This discrepancy is likely due to a reduced training duration and a manually adjusted batch size, which may have decreased intra-batch variance. Furthermore, subtle differences in post-processing pipelines across evaluation frameworks can influence FID scores, particularly due to aliasing artefacts introduced during resizing operations (Borji, 2022). As the present framework employs `torchmetrics`, which builds upon `PyTorch`, a direct comparison with the TensorFlow-based implementation used by Ho et al. (2020) is not conclusive. Nonetheless, both the FID score and the generated samples illustrated in Fig. 3.2 indicate high visual fidelity. Given the small discrepancy

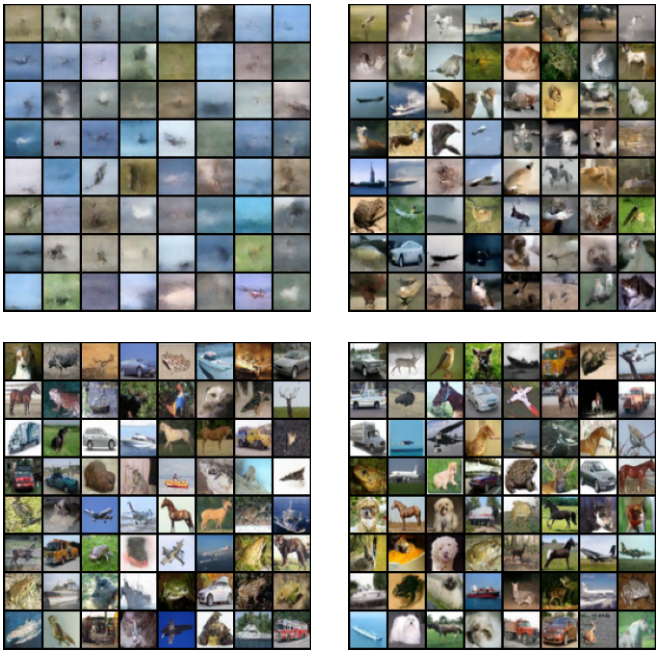


Figure 3.2: Procedural generation of artificial CIFAR-10 samples across training progression. Samples are arranged in reading order, illustrating the gradual improvement in visual fidelity as training advances. This indicates with successful diffusion model convergence.

and the strong perceptual quality, it can be concluded that the model performs effectively. This success spans the entire pipeline, from data preprocessing (including normalisation, batching, and buffering) to the model architecture and the diffusion-based training and sampling procedure. The results validate the correctness of the implementation and establish a robust foundation for future extensions to the medical imaging domain.

3.3 Medical LDM framework for 3D brain tumour segmentation

Following the successful validation of the model including data processing pipeline and parts of the evaluation process, the next step is to extend the model to 3D anomaly detection. This includes to adapt the data pipeline to volumetric medical imaging data, adapt the backbone model architecture, and shift the objective from

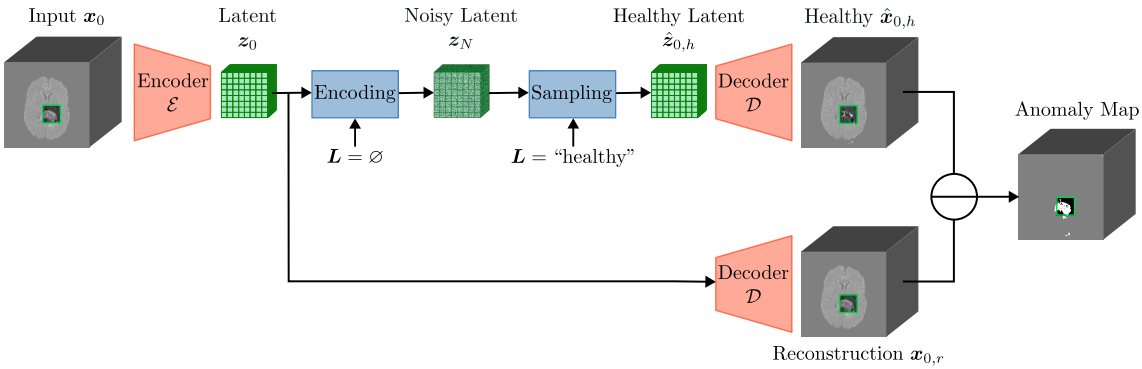


Figure 3.3: Overview of the 3D-LDM framework. The model generates a healthy counterfactual for the input patch p_{ij} (green), which is used to compute the anomaly map for segmentation. The architecture consists of a first-stage model (red), comprising an encoder $\mathcal{E}$ and decoder $\mathcal{D}$, and the DDPM (blue) operating in the latent space of the first stage. Adapted with permission from Loesch et al. (2025).

probability density estimation for sample generation to weakly-supervised brain tumour segmentation during inference.

3.3.1 Enabling 3D by transitioning to latent representations

The first part of Contribution 1 aimed to address the challenges outlined in Section 2.4.1 is the development of a novel weakly-supervised segmentation model to mitigate the computational complexity of 3D image analysis. The recently introduced LDM (Rombach et al., 2022) offers a promising solution by mapping high-dimensional input data to a lower-dimensional latent space. The standard diffusion process (Fig. 3.3, blue components) is encapsulated in a trainable first-stage model based on an autoencoder (AE) (Fig. 3.3, red components), which compresses the input data into a lower-dimensional latent space. This facilitates efficient probability density estimation with the DDPM on the compressed data. By adjusting the compression factor, determined by the number of downsampling layers (depth) of the first-stage model, the dimensionality of the input data is substantially reduced. This enables the use of computationally demanding components, such as attention mechanisms, in 3D. The information content within the latent space is controlled via regularisation, which helps balance the trade-off between information retention

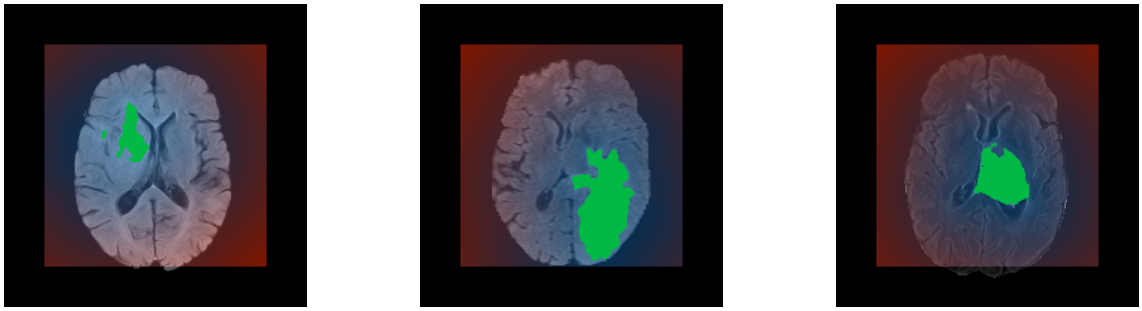


Figure 3.4: Healthy patch extraction from a diseased volume. Patches are sampled from regions defined as healthy based on their distance to the nearest lesion (green). Sampling probability increases with distance, visualised as a colour gradient from blue (low) to red (high). To avoid sampling near the volume boundaries, the probability map is offset by the patch size, indicated by black bars around the map.

and compression. However, this requires careful adaptation to ensure that essential details are not blocked by excessive regularisation or insufficient compression leaking too much information (see Section 2.3.2).

The LDM framework has demonstrated strong capabilities in high-quality medical image generation while offering substantial reductions in computational cost. These characteristics make LDMs a promising candidate for weakly-supervised anomaly detection, particularly in 3D imaging domains where efficiency constraints are prevalent. Building on these advantages, this chapter investigates the application of LDMs for weakly-supervised brain tumour segmentation in 3D MRI scans. Detailed implementation details and experimental results are provided in Section 3.4.

3.3.2 Healthy patch extraction and positional embedding

Implicitly included in Contribution 1 is addressing the challenge of extracting healthy volumetric regions from diseased individuals for 3D anomaly detection with DDPMs, which require examples of healthy anatomy to guide the model in identifying and localising anomalies (see Section 2.3.6). This work introduces a novel patch-based approach that enables the construction of pseudo-healthy datasets directly from diseased individuals. The underlying hypothesis is that, due to the typically localised nature of brain tumours, spatially distant regions remain anatomically unaffected and

can serve as valid substitutes for healthy tissue. This eliminates the need for external healthy control groups with identical MRI sequences, thereby avoiding additional data collection and ethics approvals.

Hypothesis

Tumour-induced deformation diminishes with distance, better approximating healthy brain regions.

In doing so, the method extends conventional 2D slice-based sample extraction to 3D while maintaining cohort consistency. Furthermore, the patch-based strategy substantially reduces the computational burden of 3D operations by restricting processing to small, localised subvolumes.

Detailed patch sampling procedure

Given an individual i from the brain tumor segmentation (BraTS) dataset, each MRI sequence is represented as a volume $\boldsymbol{x}_i \in \mathbb{R}^{W \times H \times D}$, where W, H, and D denote the spatial dimensions. The sampling strategy selects centroids within this volume for patches $p_i \in \mathbb{R}^{w \times h \times d}$, with $w < W, h < H, d < D$. The sampler uses a probability map that assigns each voxel a likelihood of being chosen in the current iteration. Since the anomaly-detection efficiency of DDPMs improves when trained on both healthy and diseased data (Sanchez et al., 2022), distinct sampling probabilities are assigned to healthy and diseased patch centroids.

For healthy patches, the sampling probability is determined by the distance to the nearest lesion, computed via the Euclidean distance to the surface of the binary lesion mask (see Fig. 3.4). Centroids farther are linearly assigned higher probabilities (gradient from blue to red in Fig. 3.4), reflecting the assumption that brain tumours primarily affect tissue in their vicinity, while distant regions may retain partially accurate characteristics due to brain tissue plasticity. This assumption is supported by the findings of Prasanna et al. (2019) and Subramanian et al. (2023), which demonstrate decreasing deformation severity with increasing distance to the lesion, but will be investigated separately in Section 3.4.5. To ensure no diseased patches

fall within the healthy subset, the distance map is offset by the norm of the patch dimensions. Diseased patches, on the other hand, are sampled with equal probability from the remaining regions.

The patch size $w \times h \times d$ dictates the spatial context accessible to the model during training and influences the number of patches available per subject, particularly for the healthy split. Smaller patches facilitate sampling from a broader range of healthy regions but at the expense of reduced global context available to the model during training. Conversely, larger patches provide greater spatial context but may limit the availability of healthy subvolumes without lesion contamination. Balancing these considerations is critical to ensure an optimal trade-off between healthy patch availability and the spatial context provided to the model. Given that larger spatial contexts are generally preferred in related studies (Amian & Soltaninejad, 2020; Z. Jiang et al., 2020; Y. Zhang et al., 2021), the maximum patch size is determined individually for each subject, accounting for tumour size and spatial distribution within the fixed image dimensions. For the BraTS 2023 dataset, the maximum patch size is set to 64^3 due to the following reasons:

Justification of patch size

- balancing the need for sufficient anatomical context with the requirement to sample healthy patches across a diverse cohort
- ensuring compatibility with power-of-two downsampling architectures,
- supporting multi-scale feature extraction, and
- maximising subject variability by accommodating cases with limited healthy regions.

While some individuals would permit larger patch sizes due to lesion localisation near volume edges, a generalisable approach necessitates a trade-off between context size and sampling consistency across subjects.

Patch-based sampling facilitates the efficient processing of high-resolution 3D volumes by subdividing them into localised regions. However, this approach introduces a notable limitation: ***the loss of spatial context across patches, which may***

impair the model's ability to capture intra-slice coherence and inter-patch dependencies critical for volumetric analysis. To mitigate this issue, a position embedding (PE) scheme is introduced to explicitly encode the spatial location of each patch within the volumetric structure. This encoding provides a mechanism for the model to infer the relative position of patches, thereby preserving spatial context and enhancing the representation of global anatomical structures.

This work re-utilises the timestep embedding formulation of Eq. (2.28) to encode the spatial position of each patch instead of complex embedding strategies:

$$\begin{aligned} \mathrm{PE}(pos, 2n) &= \sin\left(\frac{pos}{10000^{2n/d}}\right) \\ \mathrm{PE}(pos, 2n+1) &= \cos\left(\frac{pos}{10000^{2n/d}}\right) \end{aligned} \tag{3.1}$$

The index i represents the position within the embedding dimension d, alternating between sine and cosine functions to capture different frequency patterns. This allows the model to distinguish spatial positions across dimensions. For each patch, the position *pos* refers to its starting coordinates in the three spatial dimensions. To create the final PE, the positional encodings of these three coordinates are concatenated and passed through a learnable embedding layer. This embedding layer maps the positional information into a feature space of size d_{out}. The learnable PE help the model understand each patch's location within the volume and adjust for spatial context, even when starting positions are similar. The final positional embedding is combined with the timestep and class embeddings (see Sections 2.3.1 and 2.3.3), providing spatial and temporal information for estimating the input data's probability density.

Finally, the patch-based training strategy helps mitigate the computational cost of 3D operations. By subdividing large 3D medical volumes into smaller, localised patches, the memory and processing requirements are substantially reduced. For example, with the BraTS dataset having a spatial dimension of $240 \times 240 \times 155$, a patch size of 64^3 reduces the spatial resolution by a factor of approximately 34. This

Table 3.1: Parameter configurations used for investigating the first-stage model.

PE	None			Sinusoidal	
Codebook configuration	A	B	C	A	B
Hidden channels	64	64	128	64	64
Codebook embedding dimension	8	16	32	8	16
Codebook codes	16384	8192	16384	16384	8192
Residual blocks	1	1	4	1	1
Batch size	8				
Commitment loss weight	0.25				
Channel factor	$[2, 4]$				
Discriminator channels	64				
Discriminator layers	3				
Discriminator loss	Hinge				
Dropout rate	0.0				
EMA decay	0.9				
Learning rate	3×10^{-5}				

reduction allows for more efficient use of 3D convolutions and attention layers while maintaining manageable computational demands.

3.3.3 First-stage model investigation

The quality of the latent space is critical for the performance of LDMs, as it directly influences the model's ability to learn efficient and informative representations (Graham et al., 2023; Pinaya, Graham, et al., 2022; Rombach et al., 2022). As outlined in Section 2.3.2, 3D-autoencoder with vector quantisation regularisation (VQAE) have demonstrated strong capabilities for compressing and encoding high-dimensional medical imaging data. However, the specific configuration of the latent space, particularly the structure of the vector quantisation (VQ) layer and the number of learnable parameters, requires careful adaptation to the target task. The focus of this section is to systematically investigate how its core components can be fine-tuned for weakly-supervised brain tumour segmentation. In this context,

a 3D-VQAE with adversarial training is evaluated across different configurations to identify the optimal balance between compression efficiency and reconstruction quality. The goal is to ensure that the latent space retains essential anatomical features while minimising information loss during the encoding process. Due to the structural equivalence between the 3D-VQAE and a 3D-vector quantized generative adversarial network (VQ-GAN) model, the latter terminology is adopted throughout this chapter for consistency. The following configurations are evaluated with detailed descriptions provided in Table 3.1:

- Configuration A: VQ codebook with size of 8×16384 with 64 hidden channels and 1 residual block per layer
- Configuration B: VQ codebook with size of 16×8192 with 64 hidden channels and 1 residual block per layer
- Configuration C: VQ codebook with size of 32×16384 with 128 hidden channels and 4 residual blocks per layer

The configurations are selected to investigated the influence of the codebook size and the required complexity of the model for sufficient compression. Two key aspects of the selected configurations are particularly noteworthy: (1) the choice of the compression factor, and (2) the absence of Configuration C for sinusoidal PE. Firstly, the compression factor is a critical parameter that determines both the dimensionality of the latent space and the quality of the compressed representation. A higher compression factor results in a more compact latent space, which can lead to substantial information loss and impaired reconstruction quality. Conversely, a lower compression factor preserves more spatial and contextual information but increases the computational complexity of the model. For DDPMs, the compression factor directly influences the residual spatial dimensions in the latent space, and limits the depth of the backbone U-Net by constraining the number of layers. To balance information retention and computational feasibility, this study adopts a compression factor of 2, ensuring sufficient spatial dimensionality to facilitate deeper architectural configurations in the diffusion process given the comparatively "small" maximal patch size of 64^3. This choice is conceptually supported by Khader et al. (2023)

and Rombach et al. (2022), who found that a compression factor of 4 was sufficient for models with a spatial resolution of 256^2. Given the lower spatial resolution per dimension in this work, a compression factor of 2 mitigates computational complexity while maintaining a rich feature representation in the latent space, thus facilitating high-quality reconstructions for the subsequent diffusion process.

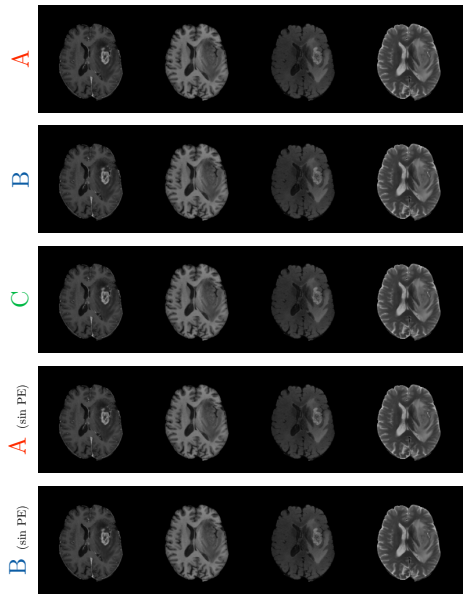
Secondly, the absence of configuration C for sinusoidal PE is due to the iterative nature of the investigation and the negligible performance increase observed in the already minimal reconstruction error without PE, as shown in Tables A.1 and A.2. As a result, the configuration was not further investigated with sinusoidal PE.

Each configuration is trained using the 80% training split of the BraTS 2023 dataset, with the remaining 20% reserved for validation and testing split equally. The test set is utilised for the final evaluation and consists of 125 subjects. To utilise the first-stage model for the subsequent LDM, the same patch size of 64^3 is utilised, with random extraction of healthy and diseased locations from the volume.

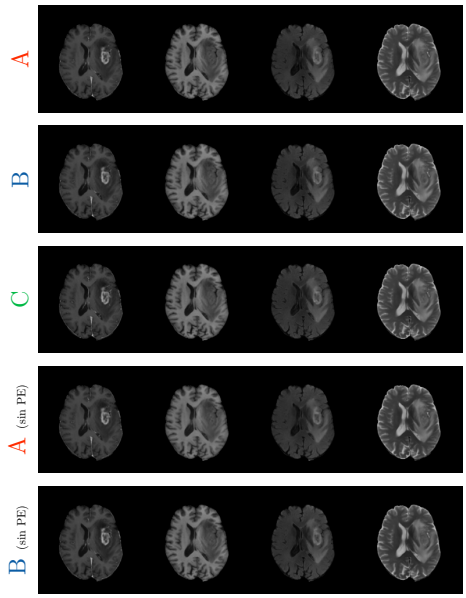
The reconstruction quality is assessed through the maximum L_1 and L_2 errors across the entire 3D volume of each individual. To investigate outliers, the 95-th and 99-th percentiles of the reconstruction errors are computed, and the corresponding volumes are visually inspected (see Fig. 3.5). To summarise the metrics across the entire test set, the mean of each metric is calculated and shown in Tables A.1 to A.2.

Codebook size

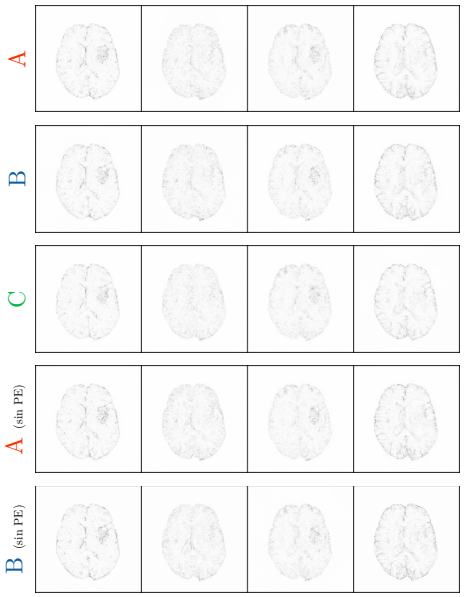
The analysis of configurations without PE indicates minimal differences in reconstruction error across MRI sequences and codebook sizes. Configuration A consistently achieves the lowest L_1 and L_2 reconstruction errors across all sequences with the exception of the L_2 error for T_2-fluid attenuated inversion recovery (T_2-FLAIR). In contrast, Configuration B, which doubles the embedding dimension while halving the number of distinct codes, exhibits a slight decrease in performance, as reflected by marginally elevated maximum L_1 and L_2 errors. Configuration C, featuring the largest codebook size and parameter count shows generally the worst performance across all tested sequences and configurations. The elevated reconstruction errors



(a) Input



(b) Reconstruction



(c) Reconstruction difference

Figure 3.5: Reconstruction differences for the first-stage model across various configurations. Columns correspond to different MRI sequences, ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. The top three rows display models without PE, while the bottom two include sinusoidal PE. (c) shows the absolute reconstruction differences: white regions indicate low reconstruction error, black regions indicate high error.

in Configuration C are likely the result of overfitting, leading to Configuration C's exclusion from further testing with PE.

The limited reconstruction error and numerical differences between each configuration are consistent with the visual results shown in Fig. 3.5. The reconstructed images exhibit high fidelity and minimal differences across configurations. The reconstruction differences appear to be the most prominent around edges with high contrast differences between pixels, which appears to be consistent across all configurations.

Positional embedding

Incorporating sinusoidal PE does not significantly affect reconstruction performance ($p > 0.05$) across sequences for either L_1 or L_2, as expected. While PE does not improve first-stage reconstruction quality, it may provide contextual information beneficial for the subsequent diffusion process in latent space. Configuration B yields slightly lower maximum errors, indicating better suppression of extreme outliers, whereas Configuration A achieves marginally lower 99th-percentile errors. These differences are minor, subject to inter-subject variability, with no clear superior configuration emerging.

Similar to the effect of the codebook, PE appears to not express any visual differences beyond the edge artefacts observed in the reconstructions. The visual results are consistent across all configurations, with no discernible differences in the quality of the reconstructed images.

In conclusion, numeric and visual results highlight that both bottleneck size and PE only marginally influence the reconstruction performance of the first-stage model in the tested configuration. However, too complex models may be prone to overfitting and highlight the requirement of a careful parameter selection. All configurations achieved good results with minimal reconstruction errors that suggest a high-quality latent space. As a result, the model with Configuration A emerged as the favourable design due to the generally superior performance, and simple and lightweight architecture with minimal parameters.

3.3.4 EDICT: alternative encoding mechanism

Contribution 1 also includes the evaluation of EDICT encoding as an alternative to conventional DDIM-based encoding for weakly-supervised brain tumour segmentation. As described in Section 2.3.6, conventional DDIM-based encoding approximates the diffusion process by assuming linear noise transitions between successive timesteps. This approximation often leads to poor reconstruction of unaltered regions and degrades under stronger conditional edits, particularly for shorter diffusion trajectories. To address these limitations, EDICT encoding has been proposed as a more robust alternative that preserves anatomical detail and supports high-fidelity visual modifications. Although EDICT has demonstrated promising results in natural image synthesis, its applicability to medical imaging, particularly in high-resolution volumetric data, remains unexplored. This work evaluates EDICT encoding within the proposed 3D-LDM framework, aiming to assess its impact on latent space quality and its ability to facilitate accurate and anatomically correct generation in the context of weakly-supervised brain tumour segmentation. Due to its potential of reducing the required number of encoding steps during the healthy counterfactual generation, EDICT addresses directly the computational demands outlined in Gap 1 in conjunction with the proposed patch-based 3D-LDM.

3.4 Experimental evaluation of the 3D-LDM

Following the detailed investigation of the first-stage model and the conceptualisation of the healthy patch extraction and EDICT encoding mechanism, the 3D-LDM is trained and evaluated on the state-of-the-art dataset for brain tumour segmentation. This constitutes the proof of concept for the 3D-LDM for weakly-supervised brain tumour segmentation. As a result, the evaluation focuses on the potential benefits of 3D operations, while mitigating computational complexity using the patch-based LDM. In addition, the evaluation investigates the EDICT encoding mechanism for efficient healthy counterfactual generation in the context of medical images for the first time.

3.4.1 Dataset, preprocessing and evaluation

The dataset for this study is provided by the 2023 BraTS challenge (see Section 2.1.1) as it contains the largest and most diverse collection of brain tumour patients. The training set contains 1251 adult study subjects, each consisting of four MRI sequences: T_1-weighted (T_1w), T_1 contrast-enhanced (T_1ce), T_2-weighted (T_2w), and T_2-FLAIR (see Section 2.1.1). The dataset is partitioned into training, validation, and testing subsets using an 80/10/10 split, respectively. Each subject is assigned to exactly one subset to ensure strict separation between splits and to avoid any data leakage during evaluation. The validation set serves to monitor overfitting and was used to find the optimal configuration (see Section 3.4.4), whereas the testing set is used to assess the generalisation performance of the model on a separate set of unseen data. Following standard procedure, each MRI sequence is normalised to $[-1, 1]$. Augmentations are restricted to horizontal flips with a probability of 50% per patch during training. Since the task involves binary segmentation, i.e. detecting tumour boundaries without distinguishing subregion composition, all tumour subregions are merged into a single binary ground truth mask (see Section 2.1.1).

Patches are obtained using the novel healthy patch sampler described in Section 3.3.2 during training and validation. The sampler allows to specify a probability for healthy and diseased patches, which was set to 50% for both respectively. The patch size was set to 64^3 due to reasons outlined in Section 3.3.2. During testing, the healthy patch sampler is replaced by a sliding-window sampler, which allows inference on the entire volume instead of random patches using a stride of 25% of the patch size.

Following standard practice in segmentation-based approaches, the Dice similarity coefficient (DSC) is used to evaluated the model's performance (see Section 2.1.2). In addition, the specificity was used or a more refined analysis of the false positive predictions, which are critical to minimise for a robust and clinically-viable model.

3.4.2 Model implementation details

As outlined in Section 3.3, the backbone U-Net architecture is inspired by the implementation of Dhariwal and Nichol (2021) and Ho et al. (2020), and includes the necessary adaptations to 3D. In addition, the novel positional embedding is added to provide locational information to the LDM as part of the patch-based training and inference process (see Section 3.3.2). The 3D-VQ-GAN validated in Section 3.3.3 serves as the first-stage model to generate compact and feature-rich latent representations. The model is pre-trained, and its parameters are fixed during the training of the second-stage DDPM. The training of the DDPM involves label embeddings and classifier-free guidance as outlined in Section 2.3.3, which facilitates the generation of a healthy counterfactual during inference. The exact hyperparameter configuration is detailed in Table 3.2 and includes the key parameters for both first- and second-stage models, collectively referred to hereafter as the ***3D-LDM***.

To enable direct comparison with state-of-the-art methods, I used a linear noise schedule with $T = 1000$ steps. LDM models typically employ fewer sampling steps than traditional DDPMs due to the reduced complexity of the latent space, facilitating faster sampling without compromising quality (Rombach et al., 2022). The model is trained using the Adam optimiser with a learning rate of 2×10^{-5} and early stopping conditioned on the validation loss to prevent overfitting. The model is trained using a single Nvidia Quadro RTX 6000 graphics card with 24 GB of memory, which imposes limitations on the model's complexity and the batch size used during training.

3.4.3 Anomaly map generation

The cornerstone of weakly-supervised segmentation with DDPMs is the computation of an anomaly map highlighting areas with non-normal characteristics. This process relies on the generation of healthy counterfactuals from the input tensor, as described in Section 2.3.6, with the classifier strength C controlling the degree of conditional guidance. Conditional sampling is applied exclusively during the generation phase, while the encoding phase uses a null embedding $\boldsymbol{L} = \varnothing$ to focus solely on anatomical

Table 3.2: Hyperparameters of the 3D-LDM, composed of the 3D-VQ-GAN and 3D-DDPM, compared against the state-of-the-art 2D-DDPM.

	Hyperparameter	**3D-LDM**	**2D-DDPM**
3D-VQ-GAN, Configuration A (fixed)	Channel factor	2,4	-
	Codebook	8×16384	
	Discriminator channels	64	
	Discriminator layers	3	
	Discriminator loss	hinge	
	Hidden channels	64	
	Pixel loss	L_1	
	Residual blocks	1	
DDPM	Activation	SiLU	
	Attention	8,16	
	Batch size	6	
	EMA decay	0.9999	
	Hidden channels	128	
	Noise schedule	linear ($T = 1000$)	
	Residual blocks	4	
	Unconditional p	0.2	
	Channel factor	1,1,2,4	1,1,2,2,4,4
	Learning rate	0.00002	0.0001
	Patch size	64^3	256^2

preservation (see Fig. 3.3) using N encoding steps. This approach ensures that the latent representation captures the full anatomical context of the input, while the generation phase allows for targeted modification towards a healthy appearance (see Section 2.3.6).

To minimise the influence of reconstruction errors introduced by the first-stage model, particularly in regions with high contrast differences (see Section 3.3.3), the anomaly map is computed using the reconstructed input tensor as the reference instead of the original input (Fig. 3.3). This ensures that discrepancies in the anomaly map reflect genuine deviations from healthy anatomy rather than artefacts of lossy compression (Section 3.4.3). Binarised predictions are derived via Otsu's thresholding

Table 3.3: Hyperparameters used in the grid search for the 3D-LDM with EDICT encoding.

Hyperparameter	**Grid Search Space**	**Description**
Number of encoding steps N	50, 100, 150, 200, 250	The encoding and decoding process uses N steps for sampling, rather than traversing the full T steps in each direction.
Classifier-free guidance scale C	0.0, 5.0, 10.0, 20.0	Strength of the classifier-free guidance mechanism (see Section 2.3.6 and Eq. (2.45))
Mixing weight ω	0.91, 0.93, 0.95, 0.97, 0.99	Weight of the stabilising layers in the EDICT mechanism (see Section 2.3.6).
Position embedding (PE)	Sinusoidal, None	PE used in the LDM for spatial information

method (Otsu, 1979). To suppress isolated outliers and enhance spatial coherence, the anomaly map is post-processed using a binary opening operation, which performs erosion followed by dilation to smooth contours and eliminate small spurious regions.

3.4.4 Optimal sampling parameter selection

The evaluation is inherently lengthy due to the iterative nature of the diffusion process, which requires encoding and decoding for each individual patch of the 3D volume during inference. To manage this complexity, a grid search was conducted on a small subset of the validation set, specifically restricted to five subjects to ensure diversity. This approach allows for a reasonable exploration of the hyperparameter space while avoiding exhaustive permutations. The validation set, being independent from the training data and therefore having no influence on the parameters of the model, provides unseen samples suitable for this task. The grid search focused on the hyperparameters described in Table 3.3 due to their direct effect on the generated healthy counterfactual, given the pre-trained 3D-LDM outlined in Section 3.4.2.

The number of encoding steps N specifically determines the quality and information content of the latent space encoding. The classifier-free guidance scale C controls the features of the generated healthy sample. Given the novel configuration of LDM with classifier-free guidance in 3D and an alternate encoding mechanism, a reasonable range for the configuration had to be determined. Dhariwal and Nichol (2021) demonstrated effective natural image manipulation using 250 steps and DDIM encoding, whereas for medical imaging, the number of encoding steps typically falls within $N \in [400, 600]$, as reported in studies on 2D models (Behrendt et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022; Wyatt et al., 2022). Given the EDICT mechanism's superior stability and efficiency but double the inference time due to the affine coupling layers (ACLs) (see Section 2.3.6), the range is adjusted to $N \in [50, 250]$ to achieve a balance between efficiency and quality. The classifier-free guidance scale C of the LDM is tested in the range $\{0.0, 5.0, 10.0, 20.0\}$ and is comparable to similar approaches without LDM (Sanchez et al., 2022), yet substantially lower than classifier-guided sampling (Wolleb et al., 2022). This range has been established during an explorative search during the development process, indicating that C beyond 20.0 severely degrades the quality of the generated healthy counterfactuals. It is important to mention that a classifier strength of $C = 0$ disables the classifier-free guidance but still allows the model to propagate label information to residual blocks using the adaptive group normalisation, as described in Sections 2.3.3 and 2.3.6. The mixing weight ω is a parameter exclusive to the EDICT mechanism and stabilises the encoding process. Given the recommended range by Wallace et al. (2023), we tested $\omega \in \{0.91, 0.93, 0.95, 0.97, 0.99\}$. Lastly, all experiments were conducted with sinusoidal PE as described in Section 3.3.2 and compared against no PE to investigate the importance of spatial information in the latent space. The choice of the PE is applied to both the first-stage and the diffusion model to ensure consistency.

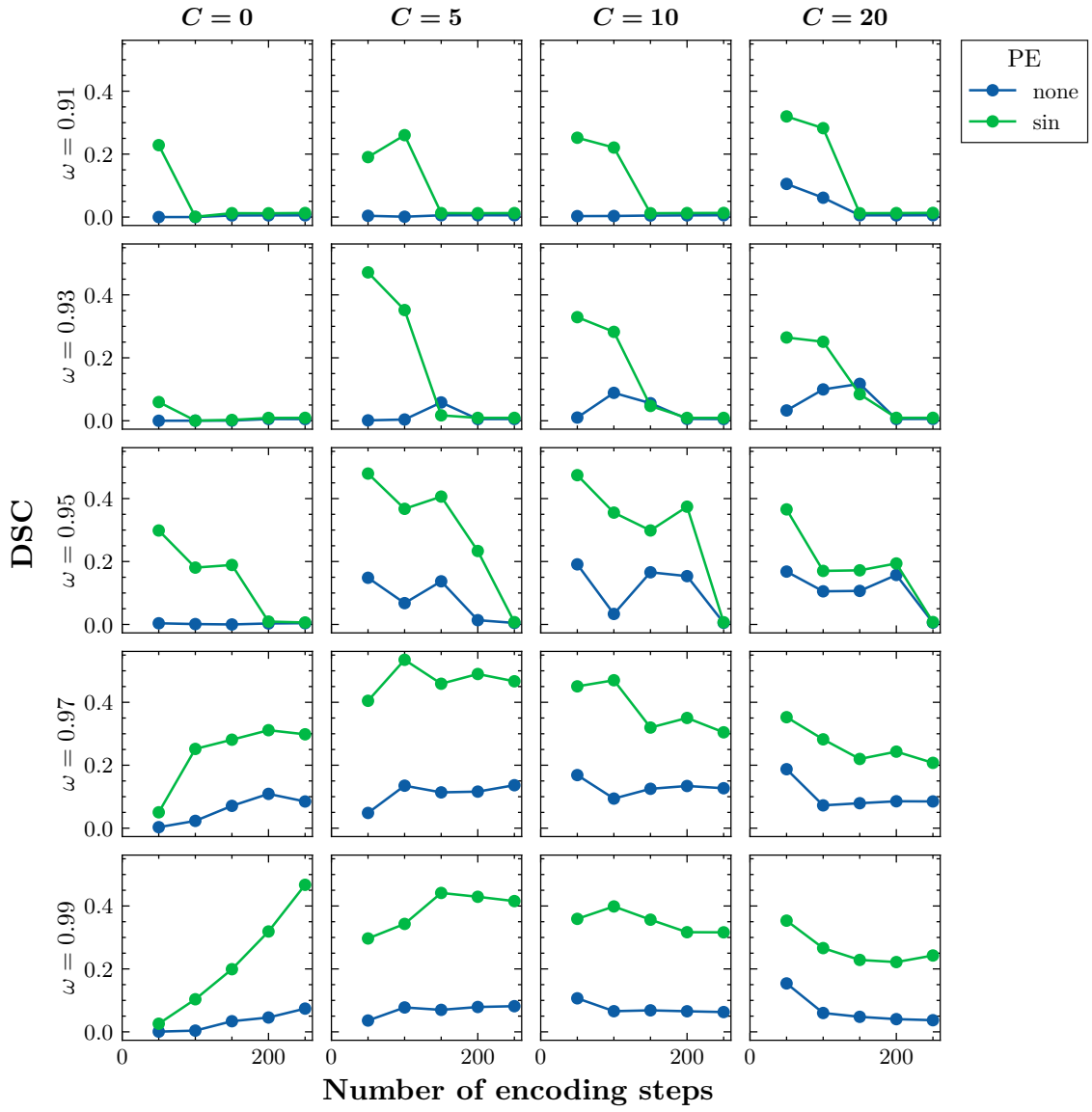


Figure 3.6: Hyperparameter grid search for the 3D-LDM. The x-axis indicates the number of encoding steps N, and the y-axis shows segmentation performance measured by the DSC. Columns correspond to different classifier-free guidance scales C, and rows represent different EDICT mixing weights ω. Line colours indicate the type of PE used in each configuration.

Findings

The results of the grid search, summarised in Fig. 3.6, demonstrate that the model's performance is highly sensitive to all tested hyperparameters. Among these, the most notable finding is the importance of PE for the patch-based 3D-LDM. Sinusoidal PE consistently outperforms models without PE, with performance improvements of up

to 40%. This underscores the critical role of spatial information in the latent space for the 3D-LDM and supports the hypothesis regarding the loss of spatial context discussed in Section 3.3.2.

Furthermore, the interaction between the number of encoding steps N and the mixing weight ω significantly impacts performance. For lower mixing weights ($\omega \in [0.91, 0.93, 0.95]$), increasing N leads to notable performance degradation, regardless of the classifier-free guidance scale C. Although this effect is less pronounced for higher mixing weights, it remains evident. These findings are partially aligned with Wallace et al. (2023), who reported optimal performance at $\omega \in 0.93, 0.97$ for $N = 50$, though our results indicate a slight shift toward higher mixing weights. Overall, larger N tends to degrade model performance, likely due to information loss from the increased noise introduced during encoding. These observations suggest that EDICT performs well for lower N when combined with higher classifier strengths but requires further evaluation on the full test set to confirm these trends.

The classifier-free guidance scale C also has a profound effect on performance. With $C = 0$, the class embedding within the residual blocks is insufficient to generate effective healthy counterfactuals, as reflected by lower DSC scores. An exception to this trend occurs at $\omega = 0.99$, where DSC scores improve for longer encoding steps ($N > 200$), possibly because the subtle counterfactual generation effects become more apparent with additional timesteps. However, in most cases, increasing C beyond $C = 10$ leads to performance deterioration due to the introduction of artefacts. These artefacts are consistently observed across all mixing weights and N, indicating that the classifier-free guidance mechanism needs to be carefully managed to balance counterfactual generation and preservation of image composition.

The optimal hyperparameter configuration was found to be $N \in [50, 150]$, $\omega = 0.97$, $C \in [5, 10]$, and sinusoidal PE. This setup yielded the best results among the tested configurations and was selected for the final evaluation. To further investigate the impact of absent classifier-guidance on artefact generation, the final evaluation also includes a configuration with $C = 0$.

3.4.5 Evaluation and comparison to state-of-the-art methods

Following the identification of the optimal hyperparameter configuration, the 3D-LDM is evaluated on an independent test set to assess its segmentation performance. A range of approaches were implemented to compare the 3D-LDM to the current state-of-the-art in weakly-supervised brain image segmentation:

1. 3D-VQ-GAN (similar to 3D-VQAE; see Section 3.3.3),
2. transformer-based 2D-WS-MTST model (H. Chen et al., 2023),
3. 2D-class activation map (CAM) model (Z. Chen et al., 2022), and
4. 2D-DDPM with classifier-free guidance and DDIM encoding (Sanchez et al., 2022; Wolleb et al., 2022).

To ensure a fair comparison between 2D and 3D models, the 2D models are trained on individual axial slices. During inference, volumetric predictions are obtained by sequentially predicting all axial slices using the 2D model.

The following sections present a series of experiments designed to evaluate key components of the proposed framework. Section 3.4.5 reports the performance of baseline methods, serving as a reference for subsequent comparisons. Section 3.4.5 explores the transition from 2D to 3D processing and its impact on model performance. The influence of EDICT encoding is examined in Section 3.4.5. Section 3.4.5 investigates the effects of conditional sampling strength on artefact formation. Section 3.4.5 analyses the role of patch overlap in mitigating edge artefacts during patch-based inference. Finally, Section 3.4.5 evaluates the hypothesis introduced in Section 3.3.2, assessing whether the proposed distance-based sampling strategy enables the use of a pseudo-healthy control.

Results of baselines methods

Baseline performances are presented first to establish a reference point, enabling a clear and robust assessment of the proposed method.

Table 3.4: Comparison of DSC scores across baseline methods and the proposed 3D-LDM configurations. Baselines include 2D-CAM[a], 2D-WS-MTST[b], 3D-VQ-GAN, and 2D-DDPM. The proposed 3D-LDM is evaluated with both DDIM and EDICT sampling strategies. The best-performing results are shown in bold.

	Model	Encoding	PE	C	N		
					50	100	150
Baseline	2D-CAM[a]	-	-	-		0.3080	
	2D-WS-MTST[b]	-	-	-		0.3494	
	3D-VQ-GAN	-	-	-		0.1863	
	2D-DDPM	DDIM	-	0	0.0946	0.1006	0.1063
				5	0.0142	0.0232	0.0364
				10	0.0160	0.0406	0.0630
Proposed	3D-LDM	DDIM	sin	0	0.0039	0.0280	0.0900
				5	0.2033	0.2119	0.2106
				10	0.2823	0.2414	0.1984
	3D-LDM	EDICT	sin	0	0.0143	0.2318	0.2312
				5	0.3614	**0.4873**	0.4025
				10	0.3827	0.3970	0.2444

[a](Z. Chen et al., 2022)
[b](H. Chen et al., 2023)

3D-VQ-GAN: The 3D-VQ-GAN achieves a DSC of 0.1863, which is consistent with values reported in the literature (Baur et al., 2021). This relatively low performance is primarily attributed to the model's ability to reconstruct abnormal regions under the specific configuration used in this experiment. As highlighted by Zimmerer et al. (2019), this behaviour is linked to the size of the latent space, which influences the model's capacity to encode relevant features. While VQAEs can achieve high anomaly detection performance with manual fine-tuning, the size of the latent space defined by the VQ in this experiment appears suboptimal.

Since the model was not explicitly tailored to the dataset and task to preserve generalisability, the latent space permits to partially reconstruct the tumour region, as illustrated in Fig. 3.7. This reduces the absolute intensity differences in areas with differing distribution to the estimate healthy distribution, and amplifies the effect of the lossy, yet minimal, compression (see Section 3.3.3). Consequently, the resulting

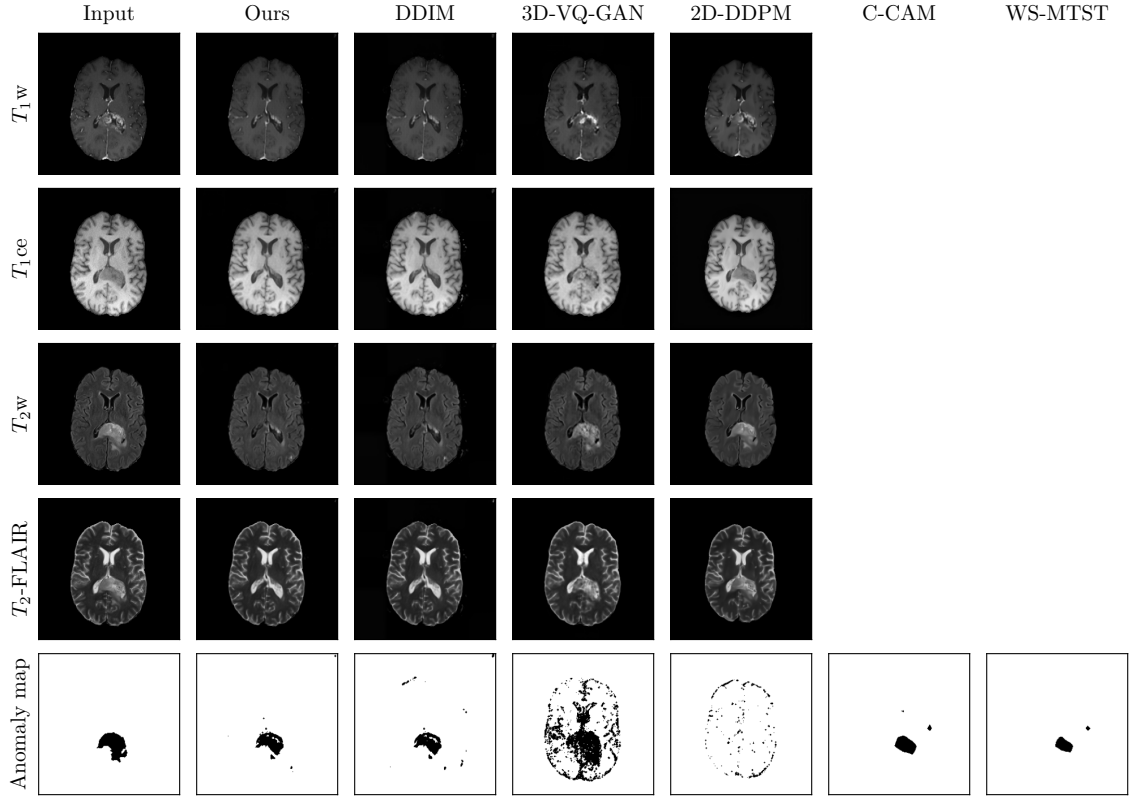


Figure 3.7: Visualisation of segmentation results for a single subject on a central slice of the volume. Columns show the input data followed by outputs from all evaluated configurations. The first four rows display the MRI sequences alongside their respective reconstructions or healthy counterfactuals, where applicable. As C-CAM and WS-MTST are not reconstruction- or generation-based models, these sequences are omitted. The final row presents the ground-truth annotation mask (first column) followed by the anomaly maps produced by each model. Adapted with permission from Loesch et al. (2025).

anomaly map exhibits high false-positive rates, capturing lesion boundaries but also falsely detecting abnormalities in healthy regions, particularly along the brain gyri.

Non-generative weakly-supervised methods: In contrast to the reconstruction-based 3D-VQ-GAN, the non-generative CAM model and its transformer-based advancement, WS-MTST, train a classifier to detect the presence of a tumour within an image, followed by the use of CAMs to visualise the spatial regions most influential for the prediction. By combining the feature maps from the final convolutional layer with the classifier's weights for the tumour class, CAMs generate heatmaps that highlight the most important areas, enhancing interpretability and

tumour boundary localisation. As a refinement of CAM, WS-MTST typically yields superior segmentation results, as evidenced by sharper tumour boundaries in Fig. 3.7. The DSCs are 0.3494 for WS-MTST and 0.3080 for CAM. However, the quality of the predictions remains constrained by the classifier's dependence on tumour feature detection, which is particularly challenging in weakly-supervised settings due to the absence of precise labels. This is further exacerbated in the presence of diffuse or small tumours, where the model may be able to estimate the core of the tumour lesion but is limited in accurately localising the boundaries. Consequently, the models exhibit high false-negative rates, missing tumour regions that are not clearly defined, as shown in Fig. 3.7.

2D-DDPM: The 2D-DDPM with classifier-free guidance and DDIM encoding achieves a best DSC of 0.1063 within the tested hyperparameter range, which is markedly lower than values reported in the literature (Pinaya, Tudosiu, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022). This discrepancy may result from (1) the reduced number of encoding steps, (2) differences in evaluation methodologies, and (3) variability in model architecture.

(1) Reduced number of encoding steps: Considering the doubling of inference time introduced by the coupling layers in EDICT encoding compared to DDIM encoding (Wallace et al., 2023), we investigated the use of a reduced number of encoding steps to balance efficiency and segmentation performance. While DDIM typically achieves optimal performance with encoding steps in the range of $N \in [400, 600]$ (Behrendt et al., 2024; Pinaya, Graham, et al., 2022; Sanchez et al., 2022; Wolleb et al., 2022; Wyatt et al., 2022), such a high step count becomes computationally infeasible for EDICT due to its increased complexity. This limitation is further compounded during 3D inference, where the processing time is directly influenced by the patch size, which is itself restricted by hardware constraints. As a result of the grid search (see Section 3.4.4), the optimal configuration for the 3D-LDM with EDICT was found to be $N \in [50, 150]$, emphasising faster inference times while preserving model performance. The feasibility of this reduction is supported by

literature indicating that EDICT can achieve substantial modifications and efficient latent space traversal with as few as $N = 50$ steps, attributed to its advanced encoding mechanism (Wallace et al., 2023). This reduction aligns with the goal of designing a segmentation model optimised for real-world clinical applications, where inference speed is critical. Additionally, the high efficiency of EDICT at low step counts provides a unique advantage, enabling practical deployment without sacrificing the model's ability to capture intricate tumour boundaries or achieve precise segmentations.

(2) Inconsistent evaluation methodologies: The evaluation methodologies employed in prior studies differ from our approach, influencing the interpretation of performance. For instance, Sanchez et al. (2022) conducted assessments on lower-resolution images, reducing false positives by allowing the model to focus on prominent features while disregarding fine details. Similarly, Wolleb et al. (2022) restricted their evaluation to centrally located slices, minimising false positives in peripheral regions with fewer foreground pixels, such as areas near the skull. Pinaya, Tudosiu, et al. (2022) trained models on a different dataset and evaluated exclusively on BraTS, limiting direct comparability.

In contrast, our approach incorporates all four MRI sequences from BraTS and evaluates performance across entire 3D volumes, increasing complexity. To ensure consistency between 2D and 3D evaluations, we applied a threshold calculated over the aggregated 3D volume rather than using slice-wise thresholds for 2D models and full-volume thresholds for 3D. While this standardisation enhances comparability, it introduces additional challenges: specifically that errors during conditional sampling in a single slice can elevate the decision boundary across the entire volume for 2D models, increasing false negative rates. In regions with pronounced intensity contrasts, these errors may simultaneously raise false positive rates, further complicating segmentation reliability. Slices without brain tissue, typically found at the superior and inferior extremes of the BraTS dataset, are particularly susceptible to such errors. Even minor deviations can lead to substantial discrepancies, shifting the binarisation threshold and impacting the final segmentation outcome.

(3) Architectural variability: The application of DDPMs in weakly-supervised learning has seen a growing number of novel publications. These methods typically modify the base model architecture by incorporating innovations such as classifier-free guidance or additional embeddings to improve performance. However, direct comparisons between these approaches are rare, as they often benchmark against non-DDPM baselines rather than one another. For classifier-guidance mechanisms, the choice of classifier strength C plays a crucial role in performance, heavily influenced by the conditioning mechanism employed - either classifier-free or classifier-guided sampling (see Section 2.3.6). For instance, while Wolleb et al. (2022) report optimal results using $C = 100$ for classifier-guided sampling, Sanchez et al. (2022) observed peak performance with $C = 3$, highlighting its dependence on the number of encoding steps N. Additionally, Sanchez et al. (2022) found that classifier-free guidance, combined with their improved conditioning mechanism, outperformed classifier-guided sampling.

Given the broad array of approaches and the absence of internal comparisons to define their relative benefits, determining the optimal configuration for each scenario remains challenging. Thus, adopting a standard configuration to obtain results and enabling consistent internal comparisons across approaches is both practical and justified. While this may not yield the best possible results for each method, it ensures uniformity in evaluation. Consequently, the baseline results reported in Section 3.4.5 retain their validity and significance in this context, serving as critical reference points for assessing improvements in segmentation performance.

Increased dimensionality

This study employed for the first time a 3D-LDM for weakly-supervised brain tumour segmentation to investigate the impact of increased dimensionality on model performance (see Contribution 1). Comparing the 3D-LDM to the 2D-DDPM reveals a substantial enhancement in performance with increasing dimensionality using the DDIM encoding mechanism. The DSC score increases by over 20% for $C > 0$, while the false positive rates do not show a consistent trend. The 2D approach occasionally

achieves better results but exhibits greater variability across configurations. In contrast, the 3D method demonstrates more stable and predictable behaviour, with less fluctuation and comparable or superior performance in the best-performing settings (Tables A.3 and 3.4).

These improvements can be attributed to two primary factors: (1) the latent space compression introduced by the LDM, and (2) the increased dimensionality of the 3D model, which effectively leverages contextual information across subvolumes.. Firstly, the latent-space embedding removes low-information, high-frequency features, enabling the subsequent DDPM layers to focus on the most relevant aspects of the data for density estimation (Rombach et al., 2022). This aligns with findings in the literature, which report comparable findings for 2D-LDM models (Pinaya, Graham, et al., 2022), given the constraints outlined above.

Secondly, the enhanced performance of the 3D-LDM also stems from its ability to integrate contextual information across subvolumes due to the increased dimensionality (Avesta et al., 2023; X. Zhou et al., 2018). This capability is further refined by the novel patch-based sampler, which implicitly constrains spatial context to regions with similar features as part of the patch-based training process. By extracting multiple subvolumes per image, this approach mitigates the common limitation of reduced inter-sample variety in 3D models, preserving diversity across training samples (Crespi et al., 2022). In addition, the 3D-LDM benefits from faster convergence, likely due to its enhanced contextual integration (Avesta et al., 2023), with the LDM compression stage further amplifying this effect.

Nevertheless, isolating the contributions of 3D operations from those of the LDM remains challenging. The LDM compression stage is not only crucial for managing hardware constraints but also supports the training of more complex DDPM architectures. Without this compression, the dimensionality and complexity of a 3D-DDPM without first-stage model would need to be substantially reduced, rendering direct comparisons with both 2D and 3D-DDPM approaches inherently biased. As discussed in Section 3.4.5, the experimental configurations were selected to balance computational feasibility and model complexity. Consequently, not all individual

configurations reported in the literature (see Section 2.3.6) were explored. Instead, a standardised configuration was adopted to ensure consistency across models, enabling a more controlled and equitable comparison of performance.

In conclusion, the integration of 3D operations, latent-space compression, and patch-based sampling markedly improves segmentation performance. The 3D-LDM consistently surpasses the 2D-DDPM in DSC and exhibits more stable specificity across configurations. It also achieves higher visual quality with effective lesion removal (Fig. 3.7) and offers improved computational efficiency through latent-space compression, resulting in substantially faster inference (Table A.4). These findings highlight the benefits of adopting a patch-based 3D-LDM for weakly-supervised brain tumour segmentation in suitable datasets, providing the first component of Contribution 1 in this research.

Effects of EDICT encoding

The second part of Contribution 1 entails the investigation of the EDICT encoding mechanism in the context of weakly-supervised brain tumour segmentation. As shown in Table 3.4, EDICT encoding consistently outperforms DDIM encoding, achieving an average DSC improvement of 15% across the tested configurations. The most pronounced enhancement is observed for $N = 100$ and $C = 5$, where the DSC reaches the maximum value of 0.4873. These results emphasise the superior performance of EDICT, particularly for shorter encoding sequences, consistent with findings reported for non-medical imaging data (Wallace et al., 2023).

A key advantage of EDICT lies in its ability to support stronger classifier-free guidance over shorter sequences, enabling more substantial alterations to image regions containing tumours. This enhanced control over the input tensor is particularly effective for reducing outliers, as evidenced by a marked reduction in false-positive predictions in background regions (see Fig. 3.7). This is reflected in an average specificity increase of 0.04 for EDICT-encoded models compared to DDIM encoding. Moderate levels of classifier-free guidance further amplify these benefits, resulting in a 25% increase in DSC relative to $C = 0$.

These findings are consistent with the results presented in Section 3.4.4, where shorter encoding sequences and moderate classifier-free guidance yielded generally superior performance. In contrast, increasing either parameter typically resulted in performance degradation for the tested configurations. A notable deviation, however, is the reduced improvement observed with $C = 0$ in this analysis, which does not increase as substantially as reported in the grid search. This discrepancy is likely attributable to the heightened complexity of the full test set, which introduces additional variability compared to the validation set.

Furthermore, the optimal classifier strength for the 3D-LDM appears to be $C = 5$, as increasing guidance scales beyond this threshold does not improve segmentation accuracy. Higher classifier strengths introduce artefacts in the foreground, leading to substantial alterations in the predicted anomaly map. This aligns with the findings of 2D-DDPM by Sanchez et al. (2022), who reported that the optimal classifier-free guidance strength is influenced by the number of encoding steps N, with $C \in [3, 5]$. However, direct comparisons remain challenging due to differences in model architectures and encoding processes.

The inference times for EDICT are approximately twice as long as those for DDIMs, as shown in Table A.4 since EDICT requires two forward passes to compute the ACLs in Eq. (2.53). Despite this increase, each subject still requires approximately 10 minutes for prediction, underscoring the critical role of LDM in enabling efficient 3D processing by managing computational complexity. Notably, the shorter encoding sequences facilitated by EDICT help mitigate prolonged inference times while achieving performance comparable to DDIM models with 400-600 steps (Sanchez et al., 2022; Wyatt et al., 2022), despite using configurations equivalent to only 200 DDIM steps.

Conditional sampling and the generation of artefacts

Despite the notable improvements in segmentation performance achieved by the 3D-LDM with EDICT encoding, the model remains susceptible to artefacts. These artefacts primarily arise from the classifier-free guidance mechanism, which can

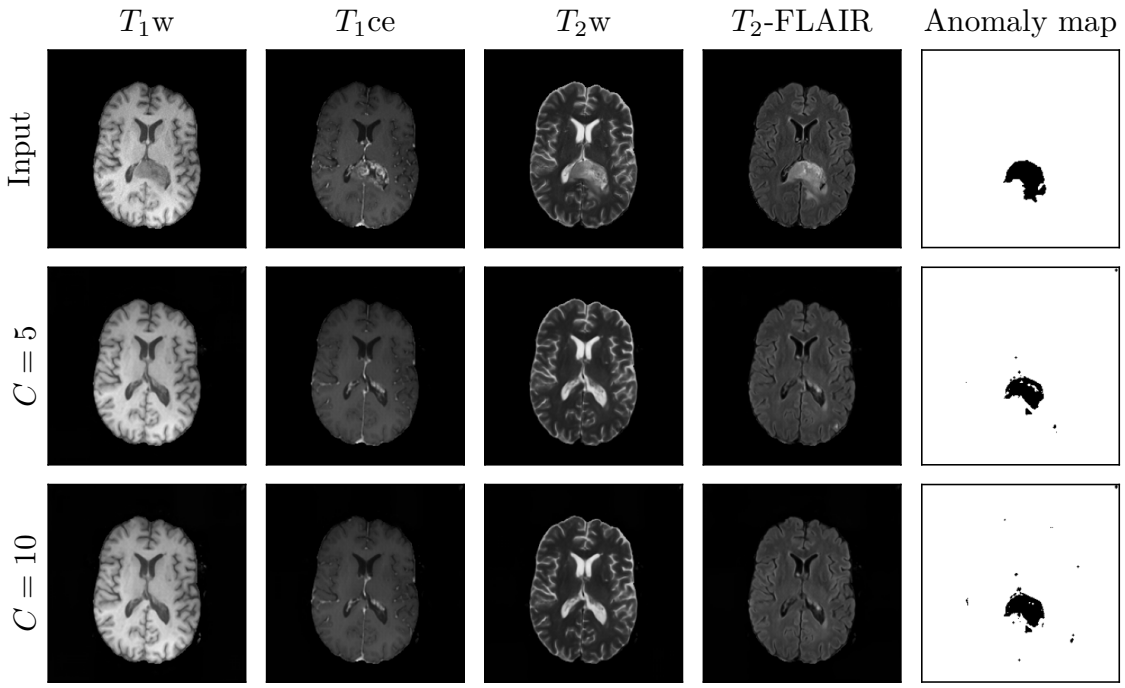


Figure 3.8: Artefacts in the 3D-LDM as a function of classifier-free guidance strength C, shown for a single subject from the BraTS 2023 dataset. Columns correspond to different MRI sequences, with the final column displaying the resulting anomaly map. The first row shows the input data, followed by the generated healthy counterfactuals for each value of C.

introduce noise and distortions into the generated samples. The severity of these artefacts is strongly correlated with the classifier-free guidance strength, with higher guidance scales exacerbating the generation of artefacts (see Fig. 3.8). This issue is particularly prominent in the background regions of the input volume, where the model fails to accurately replace the original data with its generated healthy counterpart due to insufficient anatomical context. Although lesion segmentation is improved, the presence of background artefacts can considerably impair overall performance by increasing the number of false positives.

To assess the effect of outlier predictions on model performance, an ***ablation study*** was conducted, restricting the generation of healthy counterfactuals exclusively to areas with brain tissue using the brain mask of the input data. The results, summarised in Table 3.5, reveal a substantial improvement in DSCs across all tested configurations. Notably, the 3D-LDM with DDIM encoding demonstrates substantial

Table 3.5: Relative difference in DSC from the masked ablation study.

Model	Encoding	PE	C	N		
				50	100	150
3D-LDM	DDIM	sin	0	+29	+44	+46
			5	+23	+31	+31
			10	+8	+16	+18
	EDICT	sin	0	+6	+16	+27
			5	+8	+4	+9
			10	+0	+0	+8

benefits from this masking approach, with increases up to 46%. Interestingly, under these conditions, DDIM surpasses EDICT encoding for the 3D-LDM, with a clear trend of improving DSC as N increases. In addition, the best results are now obtained without classifier-free guidance ($C = 0$) and relying solely on the adaptive group normalisation layers in the residual blocks.

Despite the increase in DSCs using the masking approach, the visual results show a slightly different picture. It is evident that the masking approach alleviates outliers outside of brain tissue and allows to complete the brain tumour lesion due to the shift in intensity thresholding. However, this comes at the cost of increasing outliers within the brain tissue, as shown in Fig. 3.9. In addition, $C = 0$ does not show a visible degree of healthification of the tumour region, despite the increase in DSC. This suggests that adaptive group normalisation layers alone are insufficient to guide the model towards generating healthy counterfactuals. As a result, tumour tissue is not effectively removed, although minor changes may still be enough to trigger the binarisation threshold.

Key findings of the masked ablation study

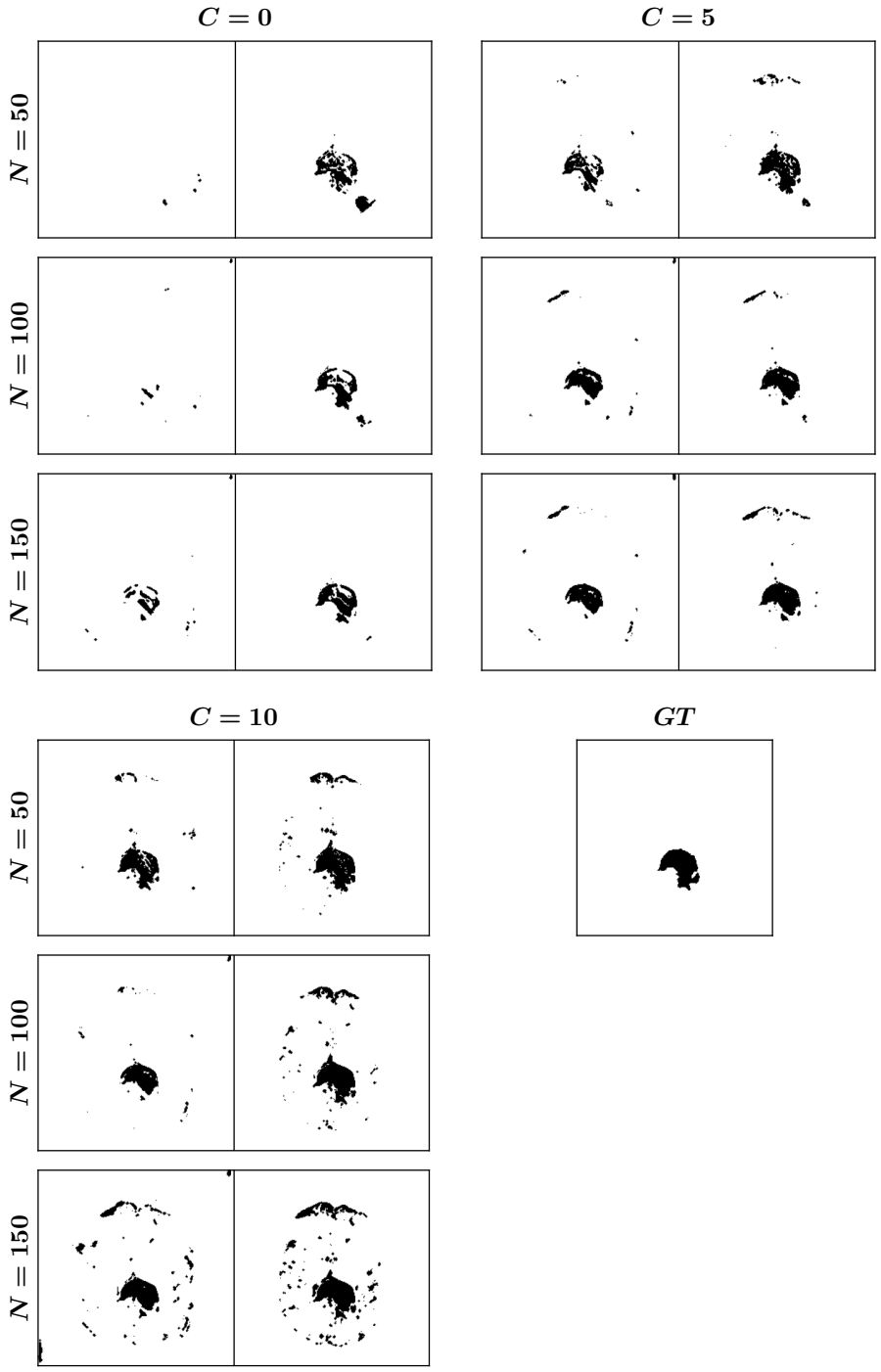
1. the importance of careful optimisation of hyperparameters,
2. the effectiveness of EDICT in mitigating outliers, and
3. the fragility of an intensity-based thresholding mechanism to calculate the anomaly map.

The interplay between the classifier strength C and the number of encoding steps N requires careful adaptation to balance performance and artefact generation. While higher classifier strengths improve segmentation accuracy under specific configurations, they also introduce artefacts in other cases, destabilising the anomaly detection framework. Similarly, the short encoding sequences facilitated faster inference but required precise tuning to balance performance and artefact generation. The comparison of the results with and without masking of the healthy counterfactual also reveals that EDICT encoding is increasingly superior to DDIM in balancing these two factors due to the substantial reduction in outliers in the background regions.

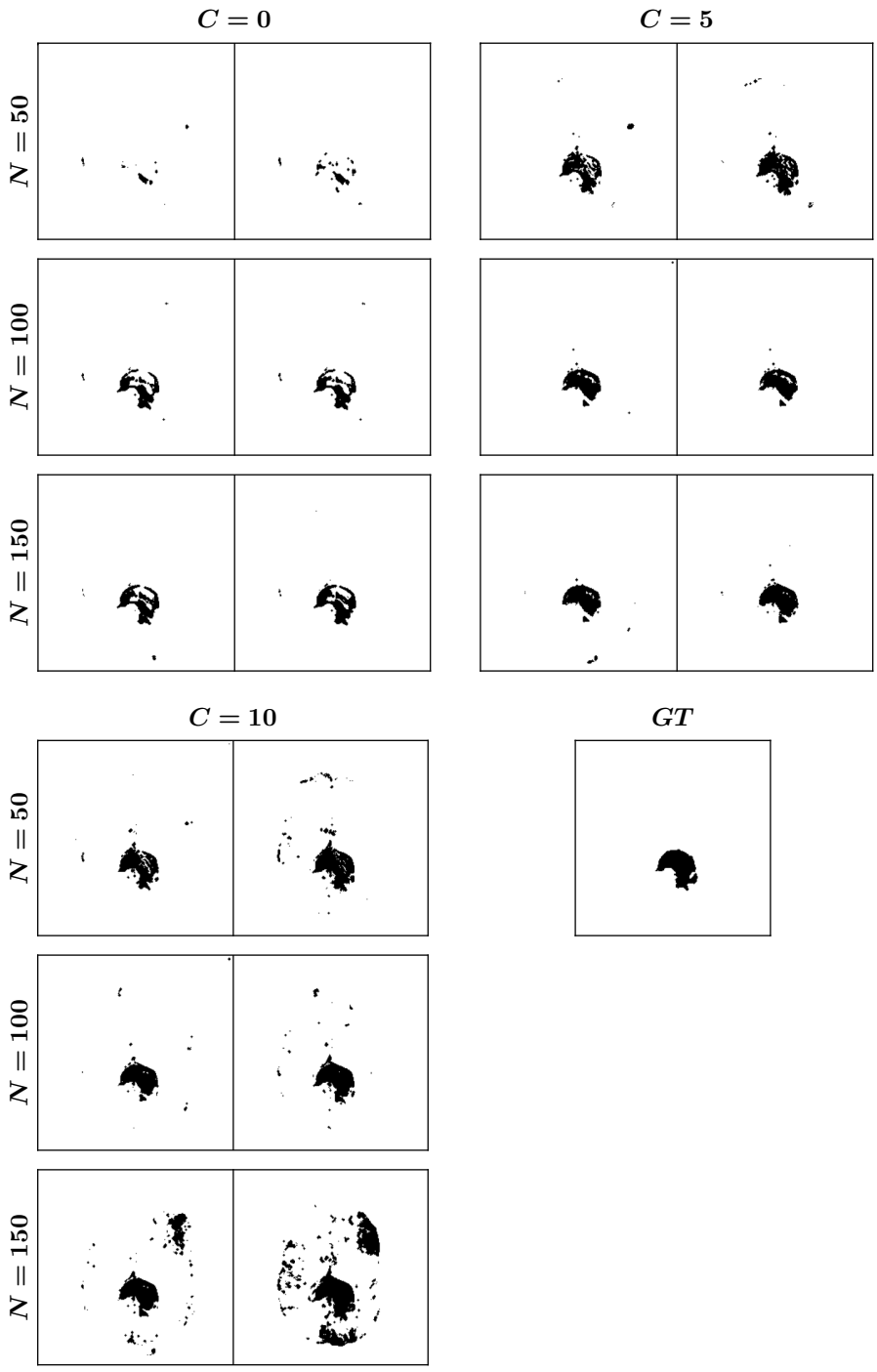
Particularly the reduction in outliers is crucial for the fragile intensity-based thresholding mechanism. Large intensity discrepancies between the input and generated samples elevate the threshold to higher values, reducing true positives in other areas, whilst increasing the prediction of false positives in these outlier regions. Specificity differences between masked and unmasked configurations underscore this limitation, reflecting the model's sensitivity to artefact-induced noise. These findings suggest to investigate more robust thresholding strategies to increase the robustness of the approach.

The effect of patch overlap

To investigate the effect of patch overlap to mitigate edge artefacts, experiments with 0%, 25% and 50% of the patch size are conducted for the best model configuration of the aforementioned experiment. This configuration utilised $N = 100$, $C = 5$ and $\omega = 0.97$ for the 3D-LDM with EDICT encoding as these were the best parameter settings found earlier. As shown in Fig. 3.10, the results indicate that larger patch overlaps improve the performance of the model, with a 50% overlap yielding the best results. However, the improvements in segmentation accuracy do not scale linearly with the computational cost. Whereas increasing the patch overlap from 0% to 25% yields a 12% improvement in DSC at the cost of a 56% increase in inference time,

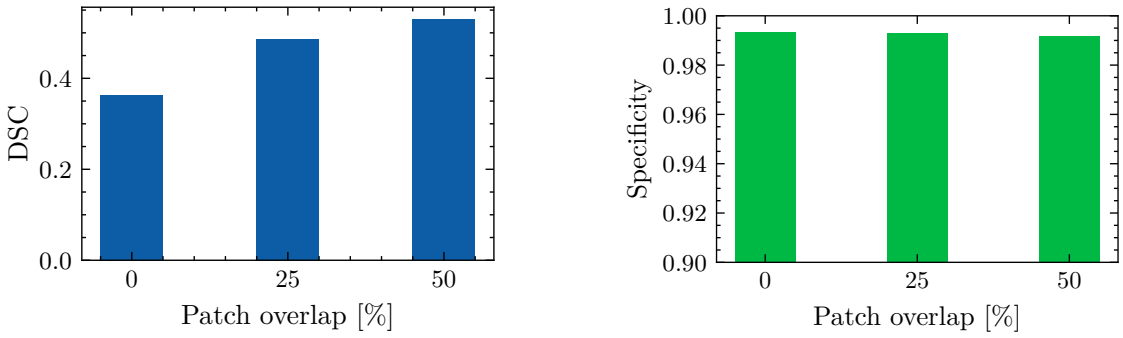


(a) DDIM



(b) EDICT

Figure 3.9: Masked anomaly maps for a single subject on a central slice of the volume. Each generated healthy counterfactual is masked using the input brain mask to suppress background outliers. Columns represent different classifier guidance strengths C, and rows correspond to varying numbers of encoding steps N. Each cell (except the ground-truth "GT") is divided into unmasked (left) and masked (right) anomaly maps.



(a) DSC (b) Specificity

Figure 3.10: Results for varying patch overlaps, expressed as a percentage of the patch size. Increased overlap improves model performance at the cost of higher computational demand (see Section 3.4.5). Results correspond to the best-performing 3D-LDM configuration reported in Table 3.4.

further increasing the overlap to 50% results in only a 4% DSC gain while inflating inference time by 160%. To balance segmentation performance with computational efficiency while ensuring sufficient coverage of volumetric permutations, I determined that a 25% patch overlap is the optimal configuration, which is used throughout the chapter, particularly for the experimental results section in Section 3.4. These results only correspond to the best-performing setting and are subject to fluctuations in improvement, whereas the increase in inference time remains consistently proportional to the overlap.

Published approaches relying on patch-based training and sliding window inference rarely test different configurations. The patch overlap percentage ranges from 10% of the patch size (C. Wang et al., 2019) to 14% (Kao et al., 2020), but is generally not a focus of the evaluation. The results of this study highlight the importance of patch overlap in balancing segmentation performance and computational efficiency, particularly in 3D models where the computational demands are markedly higher than in 2D models. The findings underscore the necessity of carefully selecting the patch overlap to ensure optimal performance while minimising computational costs, particularly in real-world clinical applications where inference speed is critical.

Table 3.6: DSC scores from the uniform sampler ablation study. The best-performing model is shown in bold.

Model	Encoding	PE	Sampling	C	N		
					50	100	150
3D-LDM	EDICT	sin	distance	0	0.0143	0.2318	0.2312
				5	0.3614	**0.4873**	0.4025
				10	0.3827	0.3970	0.2444
			uniform	0	0.0121	0.0022	0.0124
				5	0.1600	0.1440	0.1283
				10	0.2189	0.0939	0.1041

Importance of healthy sampler

To test the hypothesis from Section 3.3.2, I conducted an experiment to compare the performance of the distance-based sampler with a uniform sampler. Specifically, this experiment investigates whether distant regions from the tumour core indeed provide "healthier" tissue, which is crucial for accurately modelling the the distribution of healthy brain anatomy.

Instead of relying on a probability map related to the distance of the patch centroid to the tumour, the uniform sampler selects patches randomly from the entire volume, disregarding their proximity to the tumour. To reduce the number of experiments necessary given the rather large hyperparameter space, the experiment is conducted with the best-performing configuration of the 3D-LDM with EDICT encoding of the previous experiments. Specifically, only the sampling configurations with $N \in [50, 100, 150]$ and $C \in [0.0, 5.0, 10.0]$ were tested. With the exception of the sampling strategy, the experiment utilises the same same data processing and model architecture to facilitate an isolated comparison of the two sampling strategies.

As presented in Table 3.6, the distance-based sampler consistently outperforms the uniform strategy across all tested configurations. Improvements in DSC scores range from 10% to 35%, highlighting the effectiveness of the distance-aware sampling approach. Notably, the best-performing configuration in Table 3.4, defined by $N =$

100 and $C = 5$, yields a DSC of 0.4873 with the distance-based sampler, compared to only 0.1440 using uniform sampling. These results support the underlying hypothesis: voxels located further from the tumour core are less likely to be pathologically affected and thus more likely to reflect intact, pseudo-healthy tissue. Sampling from these regions provides "healthier" tissue, facilitating more accurate modelling of the healthy brain distribution required for artificial healthy counterfactual generation.

3.5 Limitations

While the proposed 3D-LDM framework demonstrates substantial improvements in weakly-supervised brain tumour segmentation, it is not without limitations. Artefact generation during classifier-guided inference remains a critical challenge, often manifesting in background or anatomically sparse regions and disrupting the segmentation process. These artefacts amplify the fragility of intensity-based thresholding mechanisms, shifting thresholds upward and resulting in false negatives. Specificity differences between masked and unmasked configurations further highlight the model's sensitivity to artefact-induced noise, a limitation that, despite being occasionally visible, is rarely acknowledged or addressed in previous studies.

The relationship between key parameters, including the number of encoding steps, classifier strength, and guidance mechanism, is complex. Higher classifier strengths improve accuracy in some settings but also increase the risk of artefact generation in others, which reduces the stability of anomaly detection. Shorter encoding sequences enable faster inference but require careful tuning to avoid performance degradation and preserve artefact suppression. Finally, segmentation outcomes occasionally vary between visually similar subjects, indicating limitations in model robustness and reliability.

These issues may be further exacerbated in the detection of smaller lesions, where the model's sensitivity to outliers can lead to false positives and mask subtle abnormalities. This requires an adaptation of the proposed approach to detect these abnormalities more reliably (see Chapter 4). Secondly, the observed limitations

raise the broader question of how well such models generalise; both in terms of their ability to estimate probability densities via the DDPM, and their capacity to encode fine-grained anatomical details during inference. This aspect is particularly critical in data-scarce scenarios such as paediatric imaging, where training diversity is inherently limited (see Chapter 5). Demonstrating robustness in such contexts would support the framework's adaptability and extend its applicability across clinically diverse and low-resource settings.

3.6 Conclusion

This chapter described the development and evaluation of a novel 3D-LDM framework for weakly-supervised brain tumour segmentation, addressing the first research gap of this research project (see Gap 1). Specifically, I

- developed and validated a modular DDPM framework that integrates model architecture, data handling, and evaluation protocols, demonstrating its correctness on both natural and medical image domains,
- introduced a novel 3D-LDM framework with a patch-based sampling strategy, enabling efficient and scalable processing of high-resolution volumetric data for medical image segmentation,
- demonstrated the advantages of 3D-LDMs over conventional 2D-DDPMs in the context of weakly-supervised brain tumour segmentation,
- provided the first application and evaluation of EDICT encoding in volumetric medical imaging, showing improved performance under short sampling trajectories,
- systematically analysed artefact generation and assessed the robustness of the proposed framework to outliers and conditioning noise, and
- identified and discussed key limitations of the framework, outlining directions for future research and improvement.

These efforts culminate in the following contribution:

Contribution 1

Developed a patch-based LDM for efficient 3D weakly-supervised brain tumour segmentation via anomaly detection, preserving volumetric context and integrating a robust encoding strategy to improve anatomical fidelity.

These findings directly influenced the direction of research of the following chapters, focusing on the detection of small lesions in Chapter 4 and the investigation of the generalisability of the proposed approach in Chapter 5.

Chapter 4

Super-resolution latent diffusion model for small lesion detection

Parts of this chapter have been submitted for review to "Medical Image Analysis". See Publication 2 for more details.

Building on the findings of the previous chapter, this chapter examines the capacity of weakly-supervised denoising diffusion probabilistic models (DDPMs) to detect small brain tumours. Particular attention is given to the ability of these models to capture such fine-grained abnormalities and the extent to which super-resolution (SR) influences segmentation performance. As such, it addresses the research gap identified in Section 2.4.2.

The chapter starts with an introduction and problem formulation in Section 4.1. Afterwards, the chapter is split into two main components: the synthetic generation of small brain tumours in Section 4.3 and the detection of small lesions using SR in Section 4.4. The chapter closes with a discussion of limitations of the approach in Section 4.5 and a conclusion outlining the key findings and contributions in Section 4.6.

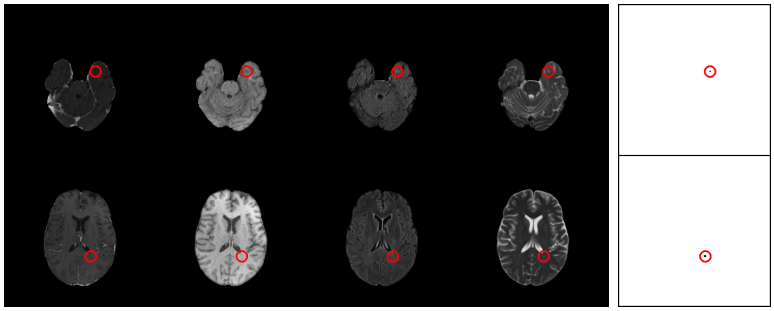


Figure 4.1: Visualisation of small lesions with a maximal diameter below 10 mm. Each row corresponds to a different subject from the BraTS dataset, displaying a representative 2D slice containing the lesion. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. The final column shows the ground-truth segmentation mask, with a red circle overlay highlighting the lesion location.

4.1 Introduction and problem formulation

As outlined in Section 2.4.2, the reliable detection of small brain tumours requires high-sensitivity imaging and precise segmentation capabilities. This challenge is illustrated in Fig. 4.1, where lesions with a maximal diameter of less than 10 mm exhibit limited visibility due to their small spatial footprint. Similarly, early-stage brain tumours are inherently small and often manifest as subtle abnormalities that are easily overlooked. Crucially, timely detection is essential to minimise long-term neurological and developmental impairments, a need that is particularly pronounced in paediatric populations (see Section 2.1.6). Despite the clinical significance, small brain tumour detection remains challenging due to anatomical variability, dependence on imaging protocols, and the scarcity of dedicated datasets. The shift towards weakly-supervised brain tumour segmentation, as proposed throughout this research project, alleviates the requirement for extensive manual annotation. However, this approach may induce reduced performance due to the absence of pixel-level supervision in the already challenging task of small lesion segmentation. To investigate the capacity of DDPMs to detect small brain tumours, this chapter is conceptualised as two distinct yet related components to address the governing research gap:

Gap 2

The role of SR in enhancing weakly-supervised DDPM-based anomaly detection remains underexplored, particularly with respect to its effect on sensitivity to small or subtle brain tumours.

These two components comprise synthetic lesion generation and SR-based detection, which are detailed in the following sections.

4.1.1 Synthetic generation of small brain tumours

The first component addresses the lack of a dedicated dataset for evaluating small lesion detection performance in brain tumours within magnetic resonance imaging (MRI) in Section 4.3. To overcome this limitation, the objective is to generate a synthetic dataset comprising MRI scans of diseased individuals with precisely controlled lesion characteristics. Synthetic generation enables systematic model evaluation under well-defined conditions, removing the reliance on extensive data acquisition and manual annotation. This approach allows exact validation of subsequent detection models and provides a controlled setting to assess the influence of increased resolution and fidelity on lesion delineation. Section 4.3 is therefore guided by the following objective:

Objective 2a: Synthetic generation of small brain tumour dataset

Generate a synthetic MRI dataset with controlled small-lesion characteristics to address the lack of dedicated benchmarks and enable systematic evaluation of model sensitivity and lesion detectability.

This section was initially intended to reproduce established methods and generate a dataset suitable for evaluating small brain tumours. However, the findings in Section 4.3 highlighted clear limitations of these approaches, prompting a detailed investigation. This encompassed the role of the conditioning mechanism (Sections 4.3.2 to 4.3.5) and ultimately informed the architectural design of the proposed SR model in Section 4.4. This investigation underpins the following contribution:

Contribution 2

Developed and validated a DDPM-based model for the controlled synthesis of size-specific brain tumour lesions in multi-sequence MRI.

4.1.2 Detection of small brain tumours using super-resolution

The second component repurposes the conditioning mechanism in Section 4.4 to increase the resolution of MRI data with the aim of improving the detection of small brain tumours. Specifically, the objective is to adapt the DDPM framework to enhance the visibility of small lesions through SR techniques, and investigate the impact of this enhancement on segmentation accuracy and robustness. The approach is hereby designed as a two-stage process, where the first stage involves the increase of spatial resolution and the removal of image artefacts, which is succeeded by the conventional anomaly detection outlined in Section 2.3.6. This section builds on the findings of Section 4.3 and is guided by the following objective:

Objective 2b: Super-resolution for small brain tumour detection

Investigate a DDPM-based SR model for enhancing resolution and fidelity in weakly-supervised anomaly detection. Evaluate its impact on sensitivity and segmentation performance, with a focus on small and subtle brain tumours.

The conditioning of DDPMs enables learnable SR of MRI scans, which is hypothesised to improve the visibility of small lesions and enhance segmentation performance. The precise role of SR in the context of weakly-supervised anomaly detection for brain tumour segmentation remains insufficiently explored. This contribution directly addresses Gap 2 by investigating the effectiveness of SR as a mechanism to increase detection sensitivity and support accurate segmentation in the challenging setting of subtle abnormalities:

Contribution 3

Conducted the first systematic investigation of DDPM-based SR in weakly-supervised anomaly detection, evaluating its impact on the detectability of small brain tumours.

4.1.3 Important remarks

Building on the findings of Chapter 3, this chapter extends the previously introduced weakly-supervised DDPM-based anomaly detection framework to the domain of small brain tumour detection. It demonstrates how the same generative DDPM backbone can be adapted to two distinct but interrelated tasks through simple modifications to the conditioning signal: synthetic lesion generation, and lesion detection through SR. This design choice highlights the versatility and modularity of DDPMs, where altering only the conditional input enables a wide range of applications within a unified framework.

Despite the findings of the preceding chapter demonstrating the advantages of 3D modelling, this investigation is conducted in 2D. The decision is primarily driven by the substantial computational cost and prolonged development times associated with 3D implementations, which would hinder efficient experimentation. Additional clinical factors further support the use of 2D. Most medical imaging data are acquired and interpreted on a slice-wise basis, and high-quality co-registered 3D volumes remain scarce (Martucci et al., 2023; Verburg & de Witt Hamer, 2021). Moreover, the 2D design aligns with Response Evaluation Criteria in Solid Tumors (RECIST) guidelines, which stipulate lesion measurements in the plane of acquisition (see Section 2.1.6). These considerations establish 2D modelling as a pragmatic foundation for exploring weakly-supervised generative methods in medical imaging, while extensions to 3D remain a feasible direction for future work, as discussed in Chapter 3.

Although exact diffusion inversion via coupled transformations (EDICT) achieved superior and faster performance in Chapter 3, this chapter uses denoising diffusion implicit model (DDIM)-based encoding and sampling in Section 4.4. This choice

isolates the effect of SR, maintains consistency with existing 2D work, and enables direct comparison. Introducing both SR and a novel encoding would confound the analysis and require additional adaptation, whereas DDIM permits evaluation of SR as a single controlled factor. To reduce artefacts near structural boundaries, masking with the brain mask constrains the reconstruction to anatomically plausible regions.

Lastly, the terms ***small lesion*** and ***small brain tumour*** are used interchangeably in this chapter, unless stated otherwise, and follow the definitions provided in Section 2.1.6.

4.2 Datasets and Preprocessing

This chapter utilises three distinct datasets, each serving a specific purpose in the investigation of small lesion detection. These datasets are summarised below:

1. ***Full-size dataset (FD)***: used for training of the generative latent diffusion model (LDM), and for training and evaluation of the lesion detection LDM.
2. ***Patch-based dataset (PD)***: used for training of the SR LDM.
3. ***Synthetic dataset (SD)***: used for evaluation of the lesion detection LDM with small lesions.

All datasets are derived from the brain tumor segmentation (BraTS) dataset, chosen for its size, diversity (1251 subjects), and the availability of ground-truth segmentations. These segmentations enable the generation of lesion masks (see Section 4.3) and support the evaluation of detection models (see Section 4.4). As outlined in Section 4.1.3, the entire approach is designed in 2D to facilitate efficient model iteration and reduced inference times. Consequently, each MRI of size $240 \times 240 \times 155$ is sampled along the axial plane, yielding 155 slices of resolution 240×240. Each slice is normalised to $[-1, 1]$ and augmented during training using horizontal flipping with a probability of $p = 0.5$. Each dataset follows an 80/10/10 split for training, validation, and testing, with the exception of the SD dataset, which is exclusively used for evaluation using all generated samples.

4.2.1 Specifics about each dataset

Full-size dataset (FD) In addition to the common preprocessing steps, each 2D slice is padded to 256×256 using background pixel values. This ensures compatibility with base-2 downsampling in the DDPM architecture.

Patch-based dataset (PD) To facilitate the training of the SR LDM, a patch-based dataset is constructed to address architectural constraints at higher resolutions. Each 240×240 slice is upsampled by a factor of 2, producing 480×480 resolution slices via bicubic interpolation. From these, patches of size 256×256 are extracted. This approach satisfies both the spatial context requirements of the LDM and the latent resolution constraints, with its rationale discussed in detail in Section 4.4.1.

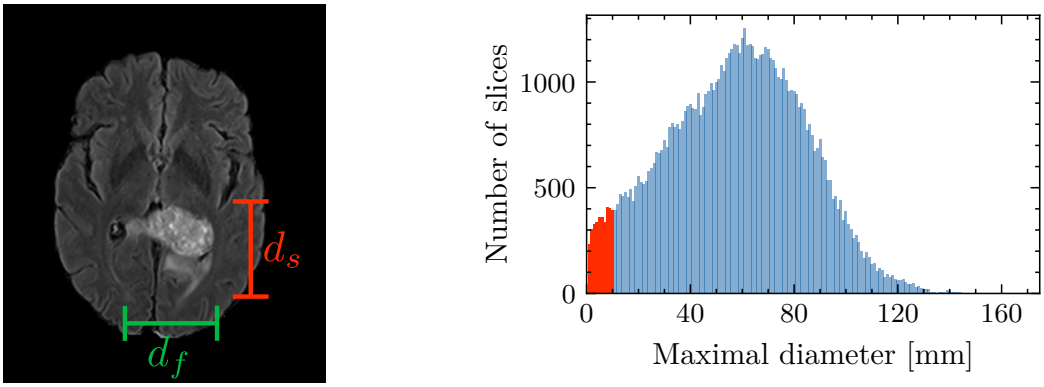
Synthetic dataset (SD) The synthetic dataset of small brain tumours is generated as described in Section 4.3.5. Since the generative 2D-LDM is trained on the FD dataset, the resulting synthetic samples are identical to the samples in FD, except for the presence of small lesions. This ensures that the synthetic dataset closely resembles real-world clinical variability, enabling the evaluation of small lesion detection performance in subsequent experiments.

4.3 Generating small brain tumours

Important remark: Lesion diameter definition

In contrast to the definition in Section 2.1.6, this chapter defines lesion diameter as the maximum extent along the two in-plane spatial axes within a given axial slice (see Fig. 4.2a). This definition is applied consistently throughout and further justified in Section 4.3.5.

To rigorously assess a model's performance in segmenting brain tumours with limited spatial extent, a dataset containing small lesions with corresponding ground-truth annotations is essential for absolute performance quantification. However,



(a) Diameter estimation (b) Number of slices per lesion diameter

Figure 4.2: Diameter calculation and distribution of lesion diameters. (a) illustrates lesion diameter estimation as the maximum extent along the sagittal (d_s) and frontal (d_f) axes within the transversal plane. (b) shows the absolute slice count per lesion diameter in the BraTS dataset, measured along the transversal plane (vertical axis). Red-shaded bars highlight lesions with a diameter below 10 mm.

the BraTS dataset includes only a limited number of axial slices depicting lesions with diameters below 10 mm (see Fig. 4.2b). With the conventional 10% test split, only 30-40 slices per lesion size would be available, which is insufficient for robust statistical analysis. This scarcity is further exacerbated by the fact that these slices often represent boundary regions of larger tumours rather than isolated small lesions, limiting their utility for evaluating segmentation performance.

Addressing the scarcity of small brain tumour samples requires either the curation of private datasets with suitable lesions or the use of generative models. The latter is more practical, as it avoids costly sample collection and annotation while ensuring a controlled evaluation environment. Furthermore, synthetic generation enables precise control over lesion characteristics, facilitating systematic assessment of segmentation performance across varying sizes and distributions, and produces data otherwise unattainable.

Building on the generative framework in Chapter 3, this section investigates how DDPMs can synthesise brain tumours of controlled size for systematic evaluation. Emphasis is placed on the conditioning mechanism, which guides generation and enables lesions with specific characteristics. Conditioning is performed using binary

lesion masks (see Section 4.3.1), a simple yet effective approach. The aim is to produce a dataset resembling clinical variability to support evaluation of small lesion detection in subsequent experiments.

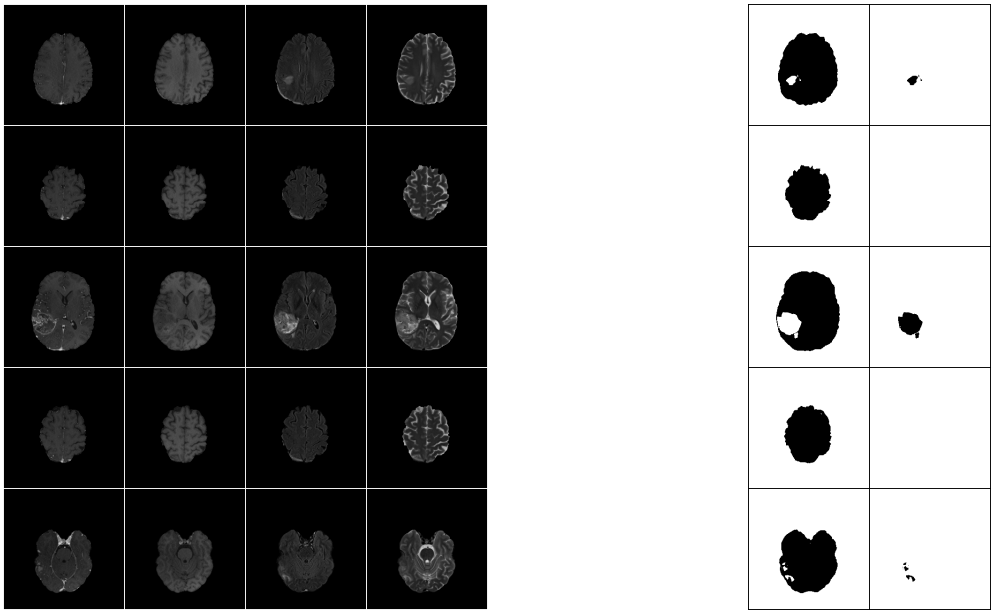
The remainder of this section is structured as follows: Section 4.3.1 introduces the conditioning signal, dataset, and evaluation strategy. Section 4.3.2 details the reproduction of Dorjsembe et al. (2024) and its limitations. Section 4.3.3 transitions to LDMs for faster, more efficient convergence (see Chapter 3) and explores conditioning strategies from Section 2.3.3. Finally, Section 4.3.5 evaluates the synthetic dataset used in Section 4.4.

4.3.1 Structural conditioning for lesion generation

To investigate the generative modelling of small brain tumours, an appropriate dataset and conditioning signal are required. This section outlines the dataset preparation, preprocessing, and evaluation approach used to enable reliable small lesion generation with DDPM. The investigated conditioning strategies are listed in Table 4.1 and visualised in Fig. B.1.

Dataset for small brain tumour generation

The generative model is trained on the FD dataset introduced in Section 4.2, which provides the required diversity and annotated cases for controlled sample synthesis. Binary lesion masks are adopted as the conditioning signal due to their simplicity and proven effectiveness in previous work (see Section 2.3.4). The multi-label ground-truth mask of the BraTS dataset is binarised into a single tumour label and combined with a binary brain mask, obtained through thresholding, to form the conditioning input (see Fig. 4.3). The brain mask provides a coarse outline of brain morphology, focusing sampling on the relevant region of interest and facilitating alignment across the input MRI sequences. Both tumour label and brain mask are one-hot encoded (see Fig. 4.3b), and supplied to the DDPM alongside the input MRI. This enables the model to learn the relationship between the both signals and allows controlled



(a) Input image (b) Conditioning mask

Figure 4.3: Example of binary lesion mask conditioning. The input image (a) and corresponding binary lesion mask (b) used for conditioning. The mask is one-hot encoded into a two-channel tensor: the first channel encodes the brain mask (left column), and the second channel encodes the optional lesion mask (right column). Modifying the lesion mask allows the model to generate images with controlled lesion characteristics and brain morphology.

generation of specific brain tumour samples by modifying lesion footprint and location within the conditioning mask during inference.

Although the use of ground-truth lesion masks in the generative process resembles supervised learning, it is important to note that this supervision is employed solely for data generation purposes. The conditioning masks are used to guide the synthesis of anatomically plausible lesions within the DDPM framework, enabling the construction of a controlled dataset. Crucially, the downstream detection model of Section 4.4 remains weakly-supervised and does not utilise any of these fine-grained lesion labels during training.

Evaluation of generative approach

The evaluation of the generative DDPM comprises both quantitative analysis and qualitative inspection for the conditioning strategies outlined in Table 4.1 and

Table 4.1: Model configurations for investigating small lesion generation with DDPMs. "Image" and "Latent" denote the feature space of the respective conditioning signal.

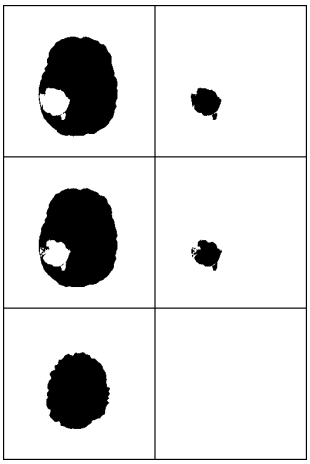
Configuration	MRI sequences	Conditioning signal	Conditioning type
Configuration G1 [a]	Image	Image	Concatenation
Configuration G2	Latent	Image	Cross-attention
Configuration G3	Latent	Latent	Concatenation
Configuration G4			Structural encoder

[a]Reproduction of Dorjsembe et al. (2024)

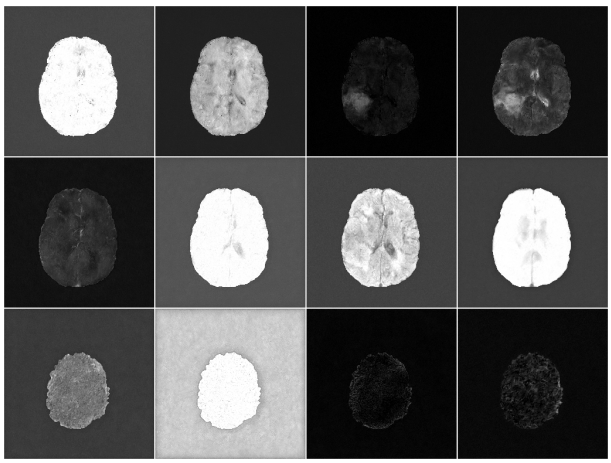
visualised in Fig. B.1. For the quantitative part, 10,000 samples were generated and assessed using a standard suite of image similarity and perceptual quality metrics. This number balances statistical robustness with the considerable computational cost of diffusion-based generation, ensuring a representative yet tractable evaluation. The chosen metrics offer complementary perspectives: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) capture pixel-level and structural similarity, multi-scale structural similarity index (MS-SSIM) extends this across multiple scales, and learned perceptual image patch similarity (LPIPS) measures perceptual similarity from deep feature activations. In addition, the Fréchet autoencoder distance (FAD) is employed, leveraging a pre-trained autoencoder (AE) trained on the BraTS dataset as a domain-specific alternative to the Inception features used in Fréchet inception distance (FID). To complement the quantitative results, representative samples are visually inspected to assess anatomical coherence, artefact presence, and the preservation of tumour-related features.

4.3.2 Concatenation of structural information

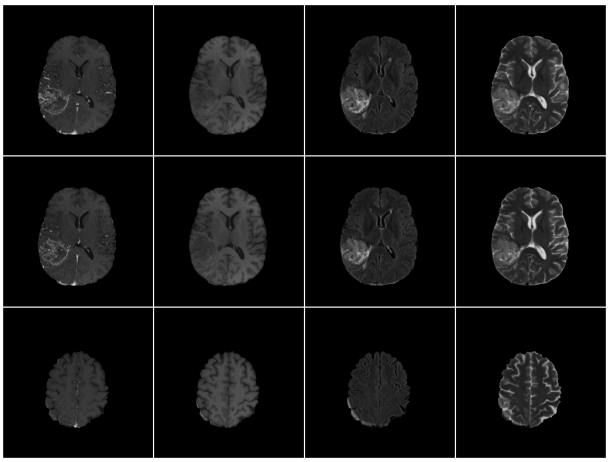
The initial method, denoted as Configuration G1, incorporates channel-wise concatenation of the one-hot encoded binary lesion mask $\boldsymbol{M}$ (described in Section 4.3.1) with the four complementary MRI sequences. Consequently, the conditioning mechanism's encoder is formulated as $\tau_\phi(\boldsymbol{M}) = \text{concat}\,[\boldsymbol{x}_t, \boldsymbol{M}]$, providing the DDPM with



(a) Conditioning



(b) Generated samples



(c) Reference sample

Figure 4.4: Generated brain images using non-latent space concatenation for conditioning. (a) shows the one-hot encoded brain mask (left column) and optional lesion target (right column) used as conditioning input. (b) displays the resulting synthetic samples, with real-world references shown in Fig. 4.4c. Rows correspond to different 2D slices from individual subjects. Columns in (b) and (c) correspond to MRI sequences ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

direct access to the conditioning signal and enabling the synthesis of lesions with distinct brain morphology and lesion characteristics. Notably, Dorjsembe et al. (2024) adapted the 2D model originally proposed by Dhariwal and Nichol (2021) for 3D data generation, employing only 250 diffusion steps. While this reduction accelerates inference, it typically increases reconstruction difficulty due to the higher per-step noise contribution. In contrast, the approach developed in this thesis retains the original architectural formulation of Dhariwal and Nichol (2021) and Rombach et al. (2022) in 2D, and extends it with a conditioning mechanism tailored to the task. To ensure high-quality sample generation and maintain stability, I decided to use a longer diffusion sequence of $T = 1000$ steps. This choice reflects a prioritisation of fidelity and robustness in training and sampling over runtime. The exact hyperparameter configuration is shown in Table B.1.

As depicted in Fig. 4.4, the generated samples exhibit poor quality and limited resemblance to real lesions. This highlights major challenges in synthesising high-quality data required for the subsequent task of detecting small brain tumours. While the generated samples roughly follow the brain shape provided as the first channel of the conditioning, they are plagued by large noise contributions, severe contrast mismatches, and an inability to resemble the human brain structure on MRI. The T_1 contrast-enhanced (T_1ce) sample (first column) in the second row of Fig. 4.4 represents an exception, as it somewhat resembles a brain structure with adequate contrast, though no lesion is visible in the desired location.

Despite attempts to improve the results by altering model complexity, prolonging training, and implementing learning rate scheduling, along with a thorough review of both my code and the publicly available code of Dorjsembe et al. (2024)[1], no notable improvements in sample quality were observed. However, this investigation highlighted several factors that may contribute to the observed results, including, but not limited to, the following:

1. (Potentially) reduced spatial context of 2D model compared to 3D model, limiting the model's ability to learn the spatial distribution of lesions across multiple planes.

2. Highly specific and unstable training process with poor sample quality as reported on the public repository[2] of Dorjsembe et al. (2024), suggesting overfitting and limited robustness of the sampling process.
3. Mismatch between model capacity and dataset composition, and (potentially) insufficient conditioning mechanism to capture the complex relationships between brain structures and lesions.

Firstly, the adaptation of the approach to a 2D model may have limited the model's ability to learn the spatial distribution of lesions across multiple planes, potentially hindering the learning process. In particular, the reduction in spatial context richness could have made it difficult for the model to capture the complex relationships between brain structures and lesions (Jung et al., 2021; Q. Zhou & Zou, 2022). However, this limitation primarily applies to the generation of 3D continuous volumes, with existing studies indicating that 2D slices still produce excellent results for various conditioning signals and datasets (Bhattacharya et al., 2025; Jung et al., 2021; Konz et al., 2024; Q. Zhou & Zou, 2022).

Secondly, the highly specific training process described by Dorjsembe et al. (2024), together with an unclear subject selection strategy, may have led to overfitting on a subset of the BraTS dataset and limited generalisation. The authors manually selected 193 subjects from the 2021 BraTS challenge but gave no details beyond choosing "high-quality images where all modalities have no distortions or artefacts" (Dorjsembe et al., 2024). This lack of clarity impedes reproducibility and raises concerns about representativeness. The subset may fail to capture the full anatomical and pathological variability of the dataset, reducing robustness on unseen cases, particularly with different imaging artefacts or anatomical variations. Reviews of their public repository[1] further report unstable training and poor sample quality, suggesting the method may not be sufficiently robust for generalised lesion generation.

In addition to these concerns, their decision to use a shortened diffusion sequence of only 250 steps is noteworthy. While this reduces training time, it substantially increases the difficulty of accurate denoising and content recovery due to larger

[1] https://github.com/mobaidoctor/med-ddpm
[2] https://github.com/mobaidoctor/med-ddpm/issues/42

noise contributions at each step. Shorter diffusion chains are known to hinder the model's ability to learn robust, high-fidelity representations, particularly in data with complex anatomical variability such as brain tumours. This choice further supports concerns of overfitting and raises doubts about robustness to unseen cases.

Lastly, although the model architecture used in this work is similar to those proposed by Dorjsembe et al. (2024) and Konz et al. (2024), it proved insufficient when applied to datasets with greater anatomical and pathological variability. Both prior works leveraged relatively constrained datasets, where limited variability may have enabled successful training despite the simplicity of the conditioning mechanism. In contrast, the 2023 BraTS dataset presents markedly more heterogeneity, which may have exceeded the model's ability to learn meaningful representations, resulting in poor generalisation and degraded sample quality. Increasing the model's capacity through deeper networks or more learnable parameters did not yield notable improvements, indicating that model complexity alone does not resolve the issue.

The findings and design choices presented by Bhattacharya et al. (2025) suggest two alternative explanations for the observed limitations. First, the conditioning mechanism itself may lack sufficient expressiveness. While their model also relies on concatenation, it employs a more advanced strategy that resembles a dedicated encoder: multi-scale features are extracted via additional learnable blocks and injected during decoding, enabling richer conditioning representations. This indicates that more sophisticated conditioning pathways may be better suited for modelling complex anatomical variability. Second, their use of downsampled input slices (128×128) substantially reduces pixel-level complexity, potentially limiting high-frequency artefacts and promoting more robust feature learning. These observations align with challenges encountered when operating directly in image space, where high-frequency, imperceptible details dominate the learning signal and can hinder the model's ability to capture relevant anatomical structures (Rombach et al., 2022). This sensitivity often leads to slower convergence and poor generalisation, particularly when training data exhibits substantial variability. To address these challenges, a promising alternative is the use of LDMs, which operate in a compressed latent space

that retains essential structural features while substantially reducing computational complexity (see Section 2.3.2 and Chapter 5). This framework offers a more robust foundation for generative tasks in complex medical domains by decoupling spatial fidelity from model capacity. As a result, the subsequent sections explore the potential of LDMs to overcome the limitations observed in image-space diffusion models.

4.3.3 Latent space conditioning

The shift to LDMs and the change in dimensionality from 3D to 2D prompts a re-evaluation of the first-stage model introduced in Section 3.3.3. In Chapter 3, this aspect received limited attention, as the focus was on developing the second-stage LDM for lesion segmentation through healthy counterfactual generation. The established autoencoder with vector quantisation regularisation (VQAE) was adopted primarily as a reliable means of encoding high-dimensional data into a tractable latent space, with its suitability for lesion generation considered secondary. In the present context, however, this transition highlights the need to reassess the appropriateness of the first-stage model and to explore potential adaptations, such as alternative regularisation schemes.

First-stage model evaluation

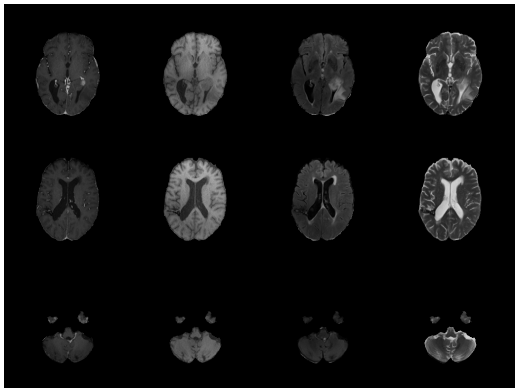
As outlined in Section 2.3.2, the first-stage model can be realised using two distinct architectures focusing on the regularisation layer: the VQAE employs a vector quantisation (VQ) layer to enforce discrete latent representations, whereas the autoencoder with Kullback-Leibler regularisation (KLAE) integrates a Kullback-Leibler (KL) divergence term, which regularises the latent distribution towards a standard normal prior. Since the subsequent diffusion model operates on learned latent representations, their fidelity and structure are critical for generating anatomically realistic lesions. Selecting an appropriate first-stage model is therefore essential to ensure that the latent space captures fine-grained lesion characteristics and supports stable training and inference.

Table 4.2: Hyperparameters of the first-stage models used for evaluating MRI sequence encoding.

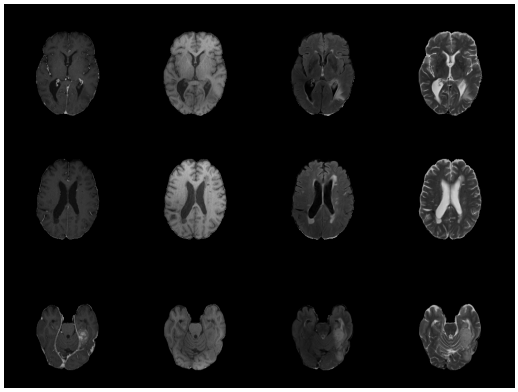
Hyperparameter	**KL-AE**	**VQ-AE**
Batch Size	4	
Channel Factor	(1, 2, 4, 4)	
Discriminator Channels	64	
Discriminator Layers	3	
Discriminator Loss	hinge	
Hidden Channels	128	
Learning Rate	3e-05	
Number of Residual Blocks	2	
Patch Size	(256, 256, 1)	
Pixel Loss	L_1	
Codebook	-	16384×4

To identify the most suitable model, this section compares VQ- and KL-regularised AEs in terms of compression quality and reconstruction accuracy, building on the analysis in Section 3.3.3. Both quantitative and qualitative evaluations are used, with particular emphasis on tumour boundary sharpness, tissue contrast, and the preservation of subtle anatomical features relevant for downstream segmentation.

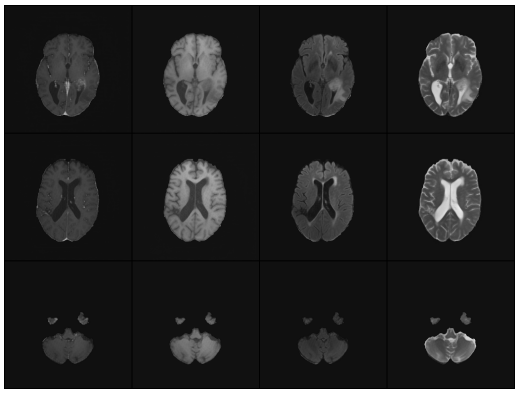
The configuration outlined in Table 4.2 was utilised to compare both fundamental regularisation techniques. Both models were trained with an additional discriminator to enhance reconstruction performance as described in Section 2.3.2. A compression factor of 8, defined by the channel factor (see Section A.1.2), was selected as the upper limit that preserves sufficient latent space dimensionality for the diffusion model. This setting balances model complexity with reconstruction fidelity and aligns with one of the recommended configurations by Rombach et al. (2022). It allows for a direct comparison of the models' maximum compression capacity, where lower reconstruction loss serves as an indicator of more effective feature preservation for downstream tasks.



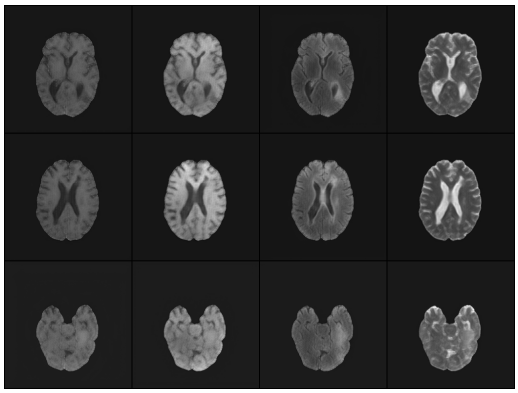
(a) KLAE: Input



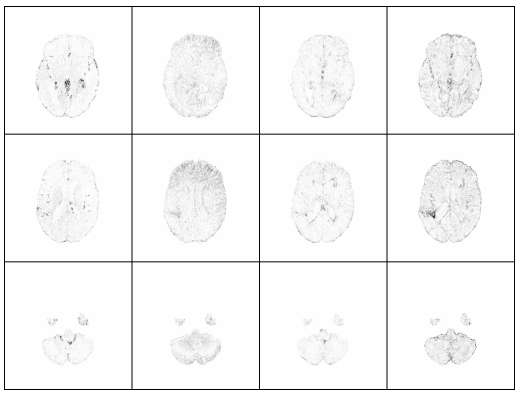
(b) VQAE: Input



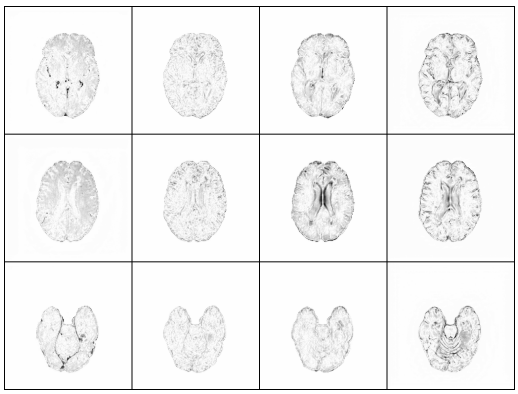
(c) KLAE: Reconstruction



(d) VQAE: Reconstruction



(e) KLAE: Difference



(f) VQAE: Difference

Figure 4.5: Comparison of first-stage models for generative LDMs. Each subplot corresponds to either the KLAE or VQAE, and includes the input image, its reconstruction, and the reconstruction difference. The latter shows the absolute reconstruction error: white regions indicate low error, black regions indicate high error. Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. Minor variations in the samples arise from differing hardware specifications during inference. The selected samples are chosen to be as visually similar as possible.

The evaluation of the first-stage models, shown in Fig. 4.5, highlights the superior reconstruction quality of the KLAE compared to the VQAE. The KL-regularised model achieves a lower average reconstruction error (0.1866 vs. 0.3242) and maintains better visual fidelity, effectively compressing the input image while preserving fine details and anatomical features. In contrast, the VQ-regularised model demonstrates limitations in capturing high-frequency details, especially lesion and brain boundary areas, resulting in a higher reconstruction error and reduced clarity. This discrepancy can be attributed to the discrete latent representation of the VQ layer, which may not fully capture the anatomical complexity.

These findings suggest that the KL-regularised AE is better suited for encoding high-dimensional image data into a compact latent space, making it the preferred choice for the first-stage model in the subsequent LDM framework.

Cross-attention conditioning

Beyond adapting the first-stage model, the transition to latent space conditioning requires changes to the conditioning mechanism. In this setting, 2D MRI sequences are encoded into a lower-dimensional latent space, while the conditioning signal (binary lesion masks) remains in image space. This spatial mismatch renders the previously employed concatenation-based conditioning approach inapplicable. To address this, a cross-attention mechanism can be employed to align conditioning inputs with differing spatial dimensionality, allowing the model to effectively incorporate image-space conditioning into the latent feature space (see Section 2.3.3). The concatenation-based approach is hereby replaced with a cross-attention layer, which is denoted as Configuration G2. The exact LDM configuration is provided in Table B.1.

As shown in Fig. B.2b, the generated samples exhibit markedly higher quality compared to the samples of the concatenation approach. High noise contributions are eliminated, and the slices present coherent anatomical structures. This improvement highlights the model's ability to capture the underlying distribution of the training data and demonstrates the benefits of latent-space conditioning. However, the generated samples show no meaningful adherence to the conditioning signal. Across

all configurations, only minor variations are observed, such as a less prominent third row in the unconditional setting, while neither the background nor lesion location aligns with the intended conditioning. Two possible causes were considered:

1. Incorrect implementation of the conditioning mechanism, and
2. Weak conditioning signal that does not sufficiently guide the generation process.

The mechanism for injecting the conditioning signal differs from the concatenation-based approach only in the use of a cross-attention layer, while the remaining implementation remains identical (see Section 4.3.2). Since the concatenation-based approach demonstrated at least partial adherence to the conditioning signal for specific cases, the first hypothesis can be rejected. This is further supported by an ablation study designed to test the influence of conditioning. Specifically, the conditioning signal was replaced with a null mask (see Fig. B.2c) and a synthetic mask (see Fig. B.2e). Across these settings, the generated samples showed only minor variations, indicating that the conditioning mechanism itself is functional but does not substantially guide the generation process. Gradient inspection further confirmed limited influence of the cross-attention module during inference, strengthening the conclusion that the second hypothesis is more likely: the single-layer cross-attention mechanism provides insufficient guidance using the binary lesion mask.

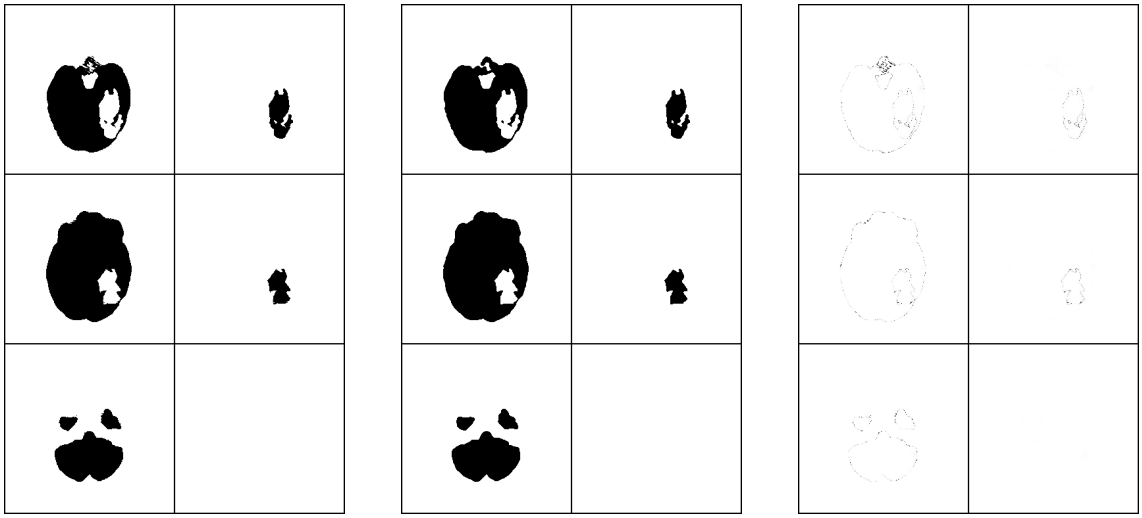
To address this limitation, an additional configuration was explored in which the binary mask was injected via cross-attention into the U-Net at multiple spatial resolutions, mimicking a spatial encoder for conditioning. However, this refinement did not improve alignment with the input mask (see Fig. B.3). To summarise, these results demonstrate that the cross-attention conditioning mechanism lacks sufficient strength to influence generation, and that even multi-scale injection fails to resolve this issue. Combined with the quadratic computational cost of attention mechanisms, these findings support the conclusion that cross-attention conditioning is both inefficient and ineffective for binary mask guidance in this setting.

Latent space conditioning for binary lesion masks

The preceding results demonstrated that the DDPM can be effectively trained in the latent space of the input MRI sequences, producing high-quality samples. The main limitation instead lies in the conditioning mechanism, where the use of a non-latent conditioning signal proved ineffective. Given the demonstrated benefits of the latent space, it is therefore natural to extend this representation to the lesion mask by encoding it into the same latent domain. This offers two key advantages: it aligns the conditioning signal with the input representation to reduce mismatches, and it provides a more expressive basis for conditioning than the sparse binary mask. The goal is to improve spatial coherence and semantic alignment of the generated samples while retaining computational efficiency.

Unlike image conditioning signals, binary lesion masks do not follow the same distribution as the MRI sequences used to train the primary first-stage model. Consequently, a separate first-stage AE is required to encode the masks into a compatible latent space. As shown in Section 4.3.3, the KLAE effectively compresses high-dimensional image data into compact latent representations. This regularisation may also benefit binary mask encoding, as the smooth latent representations reduce edge artefacts common in binary data. Such smoothing encourages the model to learn broader lesion structures rather than overfitting to pixel-level boundaries, thereby supporting more robust and generalisable lesion synthesis. Additionally, using a structurally similar first-stage model for both image data and conditioning signal further unifies the framework and enables efficient conditioning strategies that avoid the cost of cross-attention (see Section 2.3.3).

The hypothesis of an effective latent space encoding of the conditioning signal with the KLAE is supported by the results in Fig. 4.6, which evaluate the reconstruction performance of the first-stage model. This model was trained exclusively on the binary lesion masks shown in Fig. 4.4a, utilising a reduced model complexity compared to its counterpart trained on full MRI sequences (see Table B.4). Since binary masks contain inherently less structural and textural information than MRI sequences, a



(a) Input (b) Reconstruction (c) Difference

Figure 4.6: Reconstructions from the KL-regularised first-stage model trained on binary lesion masks. (a) shows the input binary masks, (b) displays the corresponding reconstructions from latent space. (c) shows the absolute reconstruction error: white regions indicate low error, black regions indicate high error Rows correspond to different 2D slices from individual subjects. Columns correspond to the two channels of the binary mask: brain mask (left) and lesion mask (right).

less complex model is sufficient to achieve high-fidelity reconstructions with limited computational resources. This is supported by the reconstructions obtained, which preserve the overall structure of both the brain and lesion and maintain spatial extent and shape while producing a smooth and continuous representation. Reconstruction errors are minimal, with only minor artefacts observed at lesion boundaries. These boundary artefacts likely stem from the model's regularisation resulting in smoothed boundaries. Nevertheless, the model demonstrates strong encoding capabilities for binary masks, reinforcing its suitability for conditioning the subsequent LDM in lesion generation.

The successful encoding of binary lesion masks in the latent space paves the way for integrating structural conditioning into the LDM framework using two additional methods: 1. direct concatenation of the latent representation of the binary mask with the latent input features (Configuration G3), and 2. utilisation of a dedicated label mask encoder to extract meaningful features adjacent to the U-Net for the given task (Configuration G4).

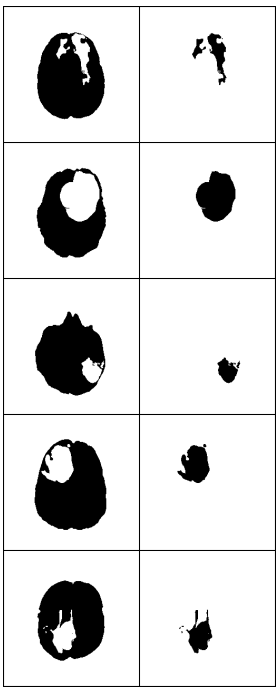
Table 4.3: Quantitative results of the generative models. The table reports scores across multiple metrics. ↓ indicates lower is better, and ↑ indicates higher is better.

Metric	Configuration G3	Configuration G4
FAD (↓)	12171.6689	12383.8115
FID (↓)	6.1738	19.1847
IS (↑)	1.0568	1.0522
LPIPS (↓)	0.0945	0.0893
MS-SSIM (↑)	0.8751	0.8820
PSNR (↑)	23.2018	23.7267
SSIM (↑)	0.8508	0.8534

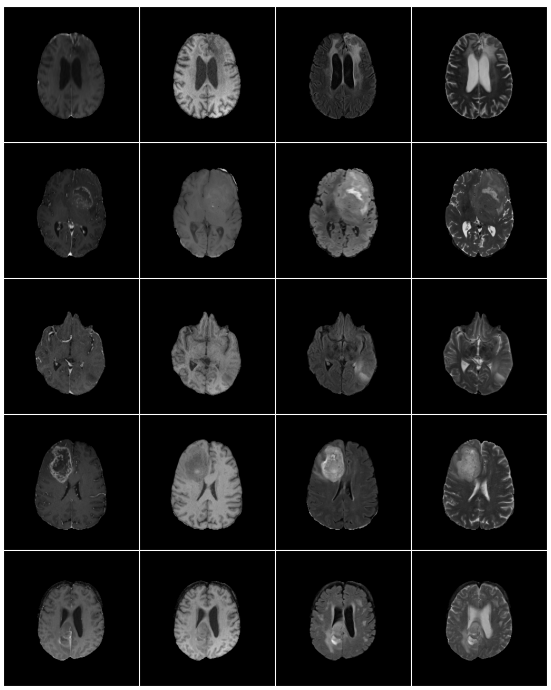
Latent space concatenation

Following the findings of the previous section, the next step is to explore the potential of latent space concatenation as a conditioning strategy (Configuration G3). It is important to note that the conditioning strategy involves ***concatenating the latent MRI sequences with the latent representation of the conditioning signal***. Although it would be theoretically viable to downsample the conditioning signal to match the latent image resolution without a dedicated first-stage model, this approach was intentionally avoided. As outlined in Section 4.3.5, interpolation during downsampling poses a substantial risk of blurring or entirely removing small lesions, which are critical for guiding the generation process. The potential loss of such features would compromise the conditioning mechanism by eliminating its ability to convey lesion-specific information, resulting in ineffective or absent guidance.

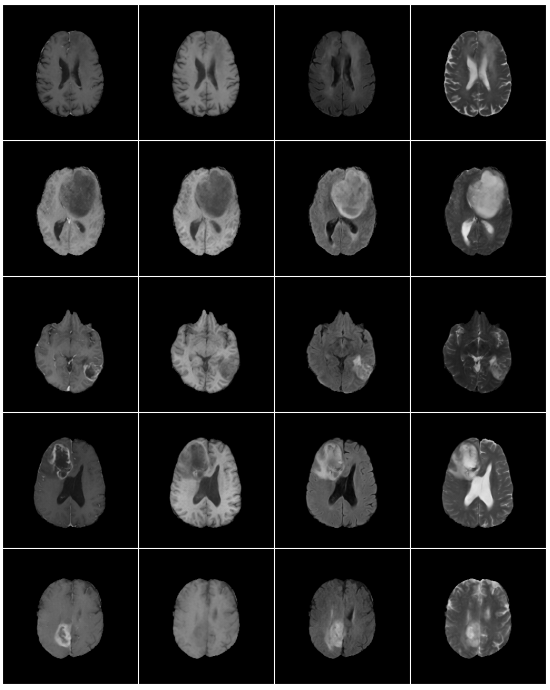
The concatenation of MRI sequences and conditioning follows the strategy outlined in Section 4.3.2, directly injecting the binary mask into the generative process (see Fig. B.1). This approach is motivated by two factors: (1) concatenation demonstrated partial success in the non-latent image space, showing adherence to the brain outline, and (2) it offers an efficient conditioning mechanism, with feature representation delegated to the subsequent U-Net layers.



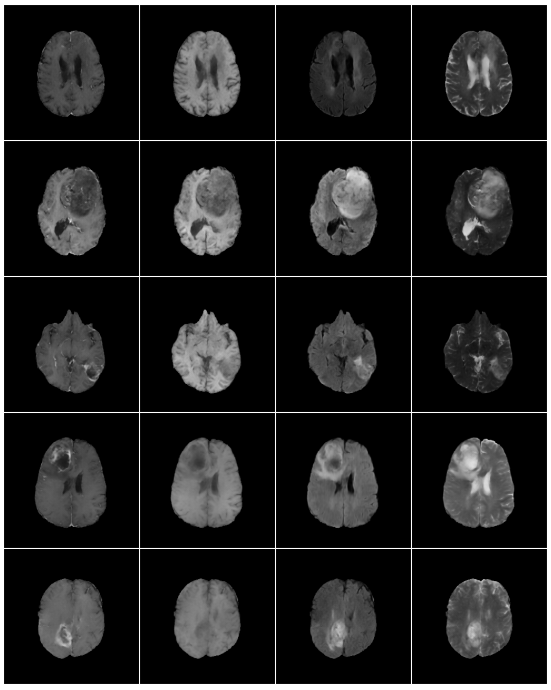
(a) Conditioning



(b) Reference sample



(c) Configuration G3: Generated



(d) Configuration G4: Generated

Figure 4.7: Generated samples for latent space conditioning. (a) shows the conditioning signal, including brain mask (first column) and lesion mask (second column). (b) displays original BraTS samples, while (c) and (d) show synthetic outputs from the two tested models. Rows correspond to different 2D slices from individual subjects. Columns in (b), (c) and (d) correspond to MRI sequences ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

Configuration G3 described in Table B.1 employs only minor parameter adjustments to accelerate training and improve convergence compared to Configuration G2, which does not have a substantial impact on the comparability. The visual results shown in Fig. 4.7c highlight the effectiveness of this strategy, with the generated samples exhibiting clear adherence to the conditioning signal. Both the brain structure and the lesion location are well-preserved, and the lesion is clearly visible at the intended spatial position. Variations between the generated samples and the reference arise from the stochastic nature of the diffusion process, combined with the relatively coarse structural guidance provided by the conditioning mask, which does not enforce precise anatomical fidelity.

These visual findings are corroborated by the quantitative evaluation presented in Table 4.3. Configuration G3 demonstrates strong generative performance, with excellent visual clarity reflected by a low FID of 6.1738 and a low LPIPS score of 0.0945. Structural consistency is similarly high, as indicated by MS-SSIM and SSIM values of 0.8751 and 0.8508, respectively, and further supported by a PSNR of 23.2018 dB. While the inception score (IS) of 1.0568 suggests relatively limited diversity, this is expected given the spatial constraints imposed by the conditioning task.

Latent space encoder

Despite its computational efficiency and the promising results demonstrated in Section 4.3.3, latent space concatenation carries the risk of weakened conditional guidance. Since the embedding is injected only at the input stage of the U-Net, its influence may become diluted over the course of the forward pass during training. This may limit the model's ability to accurately learn the complex spatial relationship between binary lesion mask and MRI sequences and could impair lesion generation performance.

An alternative approach involves using a distinct label mask encoder τ_ϕ to extract meaningful features from the binary lesion mask, which are then injected into the U-Net at various spatial levels. The encoder τ_ϕ is hereby analogous to the encoder

of the LDM backbone U-Net with specialised structural feature blocks injecting the conditional information into the U-Net at various spatial resolutions. In addition, it has been shown that this approach allows to leverage pre-trained prior knowledge of existing DDPMs by exclusively optimising the spatial encoder for the task at hand, while keeping the core of the model unaffected (see Section 2.3.3). This approach enables the efficient transfer of knowledge from a pre-trained model to a new task, considerably reducing the amount of training data, training time, and computational resources required to achieve high-quality results

To investigate the benefits of the structural condition encoder for lesion generation with pre-trained priors, the Configuration G4 configuration is employed. The first-stage model follows the setup described in Table 4.2, while the backbone U-Net uses the parameters given in Table B.1. The backbone U-Net was pre-trained for the task of weakly-supervised segmentation (see Chapter 3) with comparable complexity to Configuration G3. In addition, a secondary encoder for structural conditioning is introduced, which is structurally similar to the U-Net encoder but specifically trained on the binary lesion masks, whereas the backbone U-Net is fixed.

The visual results in Fig. 4.7d exhibit similar characteristics to those observed with latent concatenation (Configuration G3). Generated samples show clear adherence to the conditioning signal, with both brain structure and lesion location accurately preserved. Minor variations compared to the reference samples arise from the inherent stochasticity of the diffusion process and the randomness of the initial noise realisation. Overall, the results confirm that the latent encoder approach produces high-fidelity, anatomically coherent outputs while effectively integrating the structural information from the conditioning masks.

The generative metrics in Table 4.3 reveal a different trade-off. Although the FID rises to 19.1847, this metric must be interpreted cautiously. FID is suboptimal for medical images due to reliance on feature extractors pretrained on natural images and the need for large sample sizes for consistent readings (Borji, 2022; Buzuti & Thomaz, 2023; Chong & Forsyth, 2020). The LPIPS score decreases slightly to 0.0893, while MS-SSIM and SSIM increase to 0.8820 and 0.8534, respectively, with

an accompanying rise in PSNR to 23.7267 dB. These results indicate that the latent encoder improves local structural preservation and perceptual detail, even though distributional alignment appears reduced when measured by FID.

4.3.4 Conclusion of spatial conditioning investigation

The experiments demonstrate that concatenation-based conditioning (Configuration G1) fails to produce high-quality samples, as the model cannot effectively distil image-space features or align them with the conditioning signal. Transitioning to latent-space sampling confirms this, with clear improvements in sample quality. However, this shift requires an adapted conditioning mechanism due to dimensionality mismatch. Cross-attention conditioning (Configuration G2) is computationally expensive and does not provide sufficiently strong guidance.

On the other hand, both latent concatenation and latent encoder approaches yield visually coherent outputs with strong adherence to the conditioning signal. Despite its simplicity, the binary lesion mask conditioning appears sufficient for guiding the generation process. Latent concatenation (Configuration G3) offers a lightweight conditioning mechanism requiring minimal additional computation, whereas the latent encoder (Configuration G4) approach benefits from leveraging pre-trained priors of existing LDMs, which accelerates convergence and results in similar, if not slightly improved, visual and structural fidelity.

The evaluation of generative performance using FAD did not yield the expected results. Although adaptations were made compared to the method proposed by Buzuti and Thomaz (2023) to enable proper latent-space reconstruction, the resulting FAD values fall outside the range reported in the original study.

Lastly, while FID and IS serve as global quality measures, their interpretation requires caution. Both are poorly suited for medical images as they rely on models pretrained on natural images. Their reliability is further constrained by the reduced evaluation size of 10,000 images, compared to the recommended 50,000 (Borji, 2022; Chong & Forsyth, 2020), a limitation imposed by computational cost.

Given the comparable quality between both successful conditioning mechanisms and the training efficiency gained by reusing pre-trained components, Configuration G4 emerges as the preferred choice for practical applications.

4.3.5 How to generate a small lesion dataset?

Building on the generative 2D-LDM that synthesises realistic brain MRI sequences from binary lesion masks, the next step is to construct a dataset of small brain tumours using the best-performing model. By adapting the size of the lesion mask (right column of Fig. 4.3b), the model is expected to generate a synthetic dataset with exact lesion diameters. While simple shapes like ellipses or polygons could be generated semi-automatically to simulate lesions, they fail to capture the complex morphology of real tumours. To maintain clinical relevance, I decided to resize real lesions, ensuring anatomically plausible patterns and realistic spatial distributions for training and evaluation. This process and its challenges are outlined in the following section. The synthetic dataset created in this section is referred to as SD (see Section 4.2).

Binary lesion mask generation

As described in Section 2.1.1, the BraTS dataset provides binary tumour masks based on pixel-wise ground truth annotations, allowing for the rescaling of existing lesions to capture tumour characteristics. However, the application of the lesion diameter definition from Section 2.1.6, based on the maximum length within the axial slices, introduces rounding issues during the resizing process. This arises due to the limitations imposed by the isotropic voxel resolution of $1\,\mathrm{mm}^3$ in the BraTS dataset, where the specified maximum diameter often fails to align precisely with integer multiples of the pixel spacing. Consequently, resizing to an exact millimetre measure becomes impractical, resulting in inaccuracies when resizing lesions.

To address this, an ***alternative diameter definition*** is introduced, specified as the maximum spatial extent of the lesion across the remaining two anatomical axes (see Fig. 4.2a). This approach ensures precise rescaling of the lesion size within the

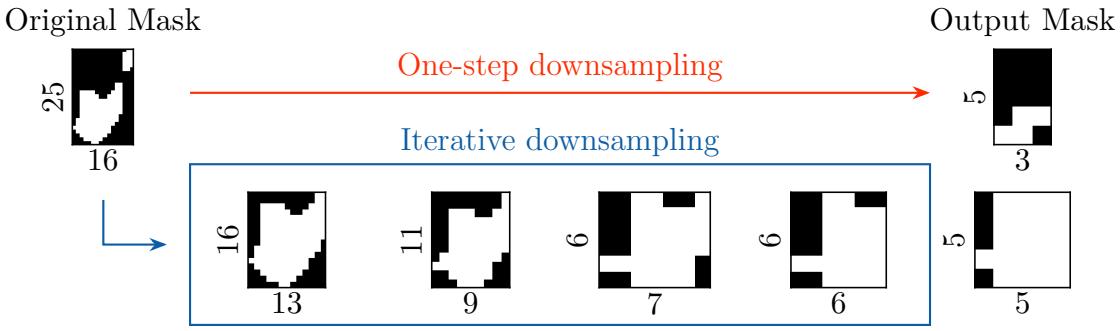


Figure 4.8: Iterative downsampling of the lesion mask to a maximum label diameter of 5 mm. Values indicate size in millimetres. The lesion bounding box is extracted and rescaled iteratively to preserve label integrity and minimise artefacts introduced by nearest-neighbour interpolation. One-step downsampling (top row), based purely on spatial scaling, results in excessive label loss, whereas iterative downsampling (bottom row) effectively maintains lesion dimensionality.

constraints of the voxel resolution, mitigating the rounding artefacts and providing a more anatomically consistent definition of lesion diameter.

Resizing is performed with nearest-neighbour interpolation to preserve the lesion's shape and structure with high fidelity. The bounding box of the lesion with dimensions $d_f \times d_s$ is extracted and rescaled to the desired size, providing accurate scaling without artefacts. The resizing factor is calculated by comparing the maximum lesion length to the desired maximum diameter d_d, with $f = d_d / \max(d_f, d_s) < 1$. However, direct resizing with nearest-neighbour interpolation can cause artefacts, especially when the lesion is small relative to the bounding box. This arises from nearest-neighbour interpolation determines the pixel value using adjacent pixels, potentially losing foreground details (see Fig. 4.8, top row). To address this, an iterative resizing approach is introduced, where a step-wise factor is applied iteratively until the lesion mask reaches the target dimensions, with $f_{i+1} = (f_i + 1)/2$. In cases where rounding prevents exact resizing, the lesion mask is cropped appropriately, ensuring accurate scaling without artefacts or loss of label information. This iterative approach guarantees that, within the limits of interpolation, the lesion preserves morphological fidelity, enabling the creation of a diverse and representative dataset of small lesions.

Creating the synthetic small brain tumour dataset

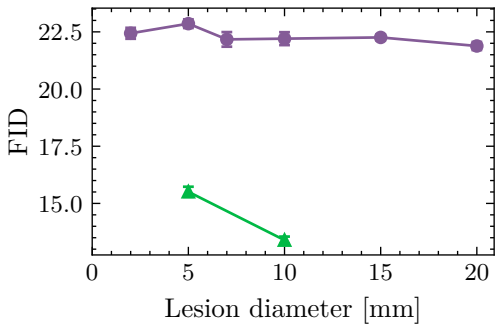
To assess generative performance, I selected 1000 subjects at random from the FD dataset and extracted five lesion-containing slices from each. The corresponding lesion masks were rescaled to the target resolution for generation using the mechanism described in Section 4.3.5. As shown in Fig. 4.9, the generative metrics remain stable across lesion sizes with only minor fluctuations for Configuration G4. These results are consistent with those obtained for real lesions in Table 4.3, suggesting that the synthetic dataset achieves comparable quality.

Visual inspection of samples in Fig. B.4 supports these findings. The conditioning mechanism performs as intended, with lesions visibly present for the untrained eye at diameters of approximately 7 mm. Additionally, the brain structure aligns well with the provided mask, indicated effective guidance through conditioning.

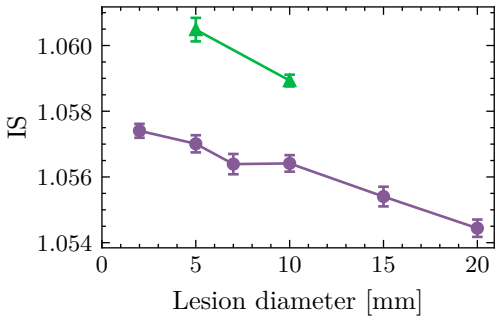
Observed distribution shift Despite the strong generative performance of Configuration G4 for both large and small-scale lesions, the synthetic dataset does not appear to follow the same distribution as the original BraTS data. This discrepancy became apparent when a pre-trained SR model was applied to the synthetic dataset in Section 4.4. The model failed to perform the expected upscaling and instead produced outputs with visible artefacts and distortions. The presence of missing regions in the upscaled outputs (second row of Fig. B.18) indicates a distribution shift in the synthetic dataset that undermines the compatibility with upscaling models trained on real data.

There are three potential strategies to mitigate this distribution shift:

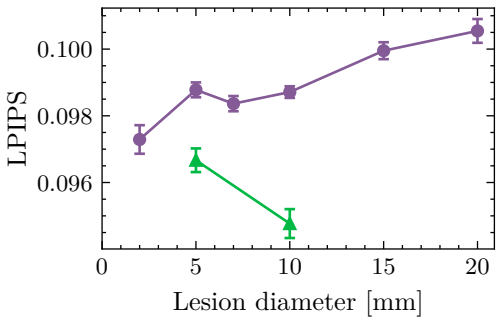
1. increase the capacity of the generative model to better capture the intricate features of the real data distribution,
2. fine-tune the second-stage upscaling model on the synthetically generated dataset, or
3. perform the evaluation on the original BraTS data without relying on exact lesion diameters.



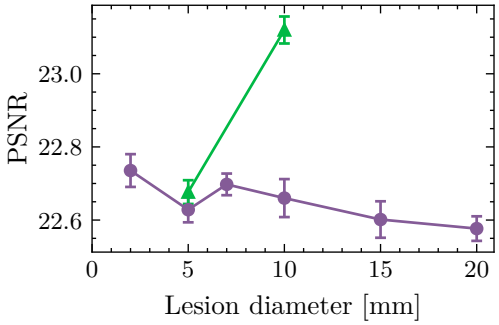
(a) FID (↓)



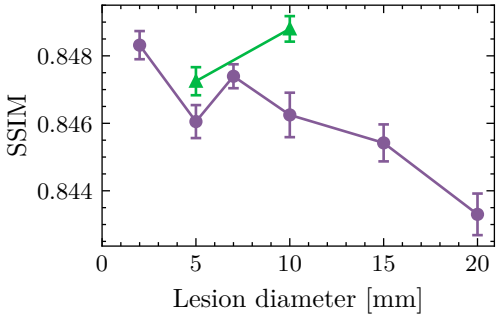
(b) IS (↑)



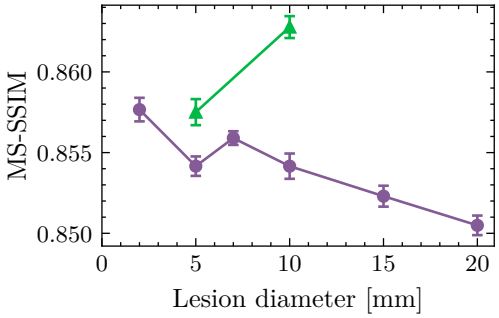
(c) LPIPS (↓)



(d) PSNR (↑)



(e) SSIM (↑)



(f) MS-SSIM (↑)

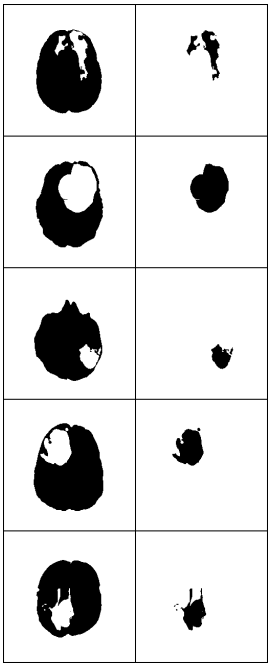
Figure 4.9: Generative metrics for the synthetic small brain tumour dataset as a function of lesion size. Results are shown for Configuration G4 (circular markers) and Configuration G3wide (triangular markers). The reduced number of data points for Configuration G3wide is due to sampling limitations during dataset creation, as described in Section 4.3.5.

This section focuses on the first approach, as improving the fidelity of the synthetic data is foundational to ensuring broader applicability. High-quality, realistic data not only facilitates fair evaluation but also enables downstream models to generalise more effectively without requiring adaptation to artefact-prone inputs or those with shifted distribution.

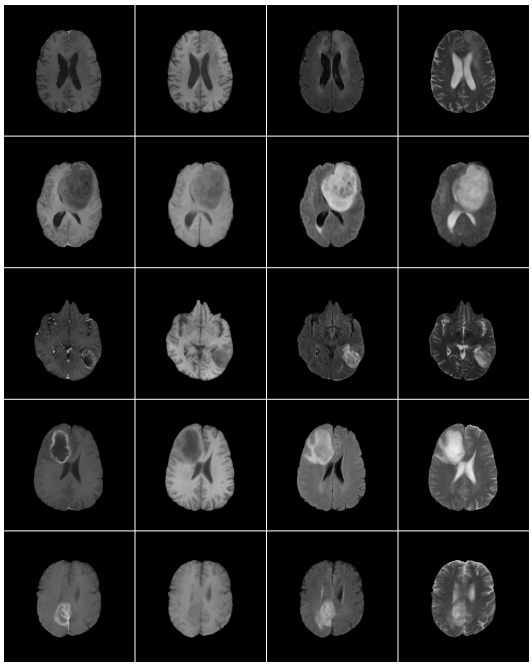
Given that latent-space concatenation and encoder-based guidance demonstrated comparable generative performance in the previous section, I investigated distribution shift using the concatenation strategy (Configuration G3). This approach was selected due to time constraints toward the end of the project and the practical advantage of requiring only a single model to be trained. In contrast, the encoder-based strategy would have necessitated a dual-model setup involving both a pre-trained generative model and a structural conditioning encoder, substantially increasing complexity and training time.

While increasing model complexity by adding layers to the U-Net backbone would theoretically enhance its ability to capture more intricate features, this approach substantially increased computational demands and caused challenges during optimisation. Instead, I expanded the capacity of the existing architecture by widening the layers, increasing the number of feature channels per layer. This adjustment preserved computational feasibility and resulted in a more stable training and optimisation process. The respective model is referred to as Configuration G3wide and is described in Table B.1.

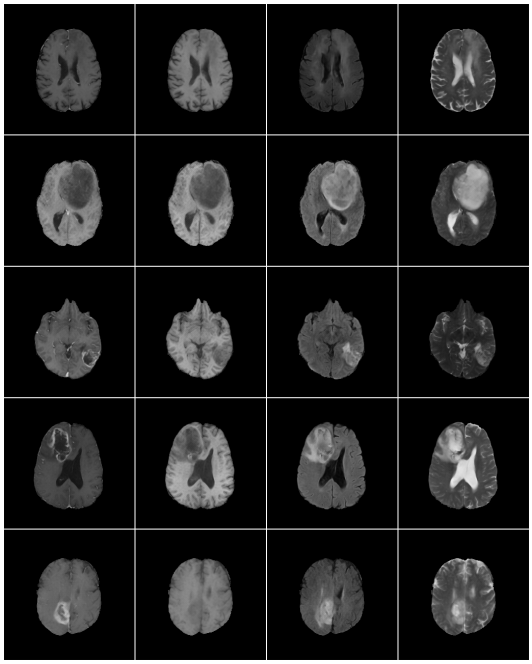
It is important to note that the evaluation was performed in the same way for real-world data as described in Section 4.3.1, whereas the synthetic lesion sizes were restricted to $d_d \in 5, 10$mm to accelerate the process and enable progression to the detection stage in Section 2.1.6. Evaluating a broader range of synthetic lesion configurations during detection was deemed infeasible due to the high computational cost of inference, as outlined in Section 4.4.4. This constraint also explains the reduced number of data points for Configuration G3wide in Fig. 4.9.



(a) Conditioning



(b) Configuration G3wide: Generated



(c) Configuration G3: Generated

Figure 4.10: Generated samples from the latent concatenation model with increased capacity (Configuration G3wide). (a) shows the conditioning signal, including brain mask (first column) and lesion mask (second column). (b) displays the resulting synthetic outputs of Configuration G3wide. (c) depicts the outputs from the lower-capacity model (Configuration G3) for direct comparison. Rows correspond to different 2D slices from individual subjects. Columns in (b) and (c) correspond to MRI sequences ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

Table 4.4: Comparison of the generative models Configuration G3 and Configuration G3wide, illustrating the effect of increased parameter count on performance. The table reports scores across a range of generative evaluation metrics. ↓ indicates lower is better, and ↑ indicates higher is better. FAD is omitted for Configuration G3wide as noted in Section 4.3.4.

Metric	Configuration G3	Configuration G3wide
FAD (↓)	12171.6689	-
FID (↓)	6.1738	6.6808
IS (↑)	1.0568	1.0564
LPIPS (↓)	0.0945	0.0755
MS-SSIM (↑)	0.8751	0.9064
PSNR (↑)	23.2018	24.6965
SSIM (↑)	0.8508	0.8659

As shown in Table 4.4, increasing the model capacity improves image quality for real lesions across several generative metrics. Specifically, LPIPS decreases to 0.0755, MS-SSIM increases to 0.9064, and PSNR rises to 24.6965 dB. While the FID slightly worsens and the IS remains unchanged, visual results in Fig. 4.10b support these findings, showing sharper and more anatomically coherent structures. Improvements are especially evident in lesion regions, where boundary definition and realism are enhanced.

The evaluation of synthetic lesions in Fig. 4.9 reveals a consistent quality increase across all tested lesion sizes and metrics. Notably, the FID improves substantially compared to previous configurations (Configuration G3 vs. Configuration G4 in Table 4.3). Gains in other metrics mirror those observed for real lesions, confirming the benefit of increased model capacity for generating high-fidelity synthetic datasets. Importantly, the severe degradation effects observed in Fig. B.18 are no longer present when applying the pre-trained SR model to the new synthetic dataset, indicating that the distribution shift has been mitigated.

4.4 Detecting small brain tumours

As mentioned in Section 2.1.6, early-stage brain tumours often manifest as small, subtle lesions, making prompt and accurate identification essential for initiating effective treatment. This is particularly crucial in paediatric populations, where early intervention can substantially mitigate the risk of irreversible neurological and developmental impairments.

In this context, the following section investigates how spatial resolution influences detectability of small lesions in MRIs. Specifically, a LDM for SR is introduced, using the same conditioning strategies as in Section 4.3 and a conditioning signal adapted to the target task.

The section is structured as follows: Section 4.4.1 outlines the experimental setup, detailing the dataset, conditioning signal composition, and training and inference process. Section 4.4.2 reports results on the original BraTS lesions and on synthetically generated small tumours. The section also includes two ablation studies to validate individual pipeline components in Section 4.4.3. Section 4.4.4 discusses inference times and computational requirements. The section concludes with a conceptual outlook on unifying the sequential approach in Section 4.4.5, followed by an overarching discussion of limitations of both generative and detection components in Section 4.5.

4.4.1 Experimental setup

The SR approach closely follows the generative framework of the previous section but differs in the conditioning signal. Instead of a binary lesion mask, the conditioning is provided by the corresponding low-resolution (LR) image, which serves as the reference for the SR task (see Section 2.3.5). Due to time constraints, only the encoder-based strategy (Configuration S4) was investigated. This choice is supported by its effectiveness in the generative experiments, promising results in preliminary tests, re-usability of pre-trained models, and demonstrated potential in natural image domains (see Section 2.3.5).

Datasets for small lesion detection

The super-resolution model is trained on the PD dataset described in Section 4.2. This dataset preserves clinical diversity while enabling patch-based training at higher resolutions, thereby supporting the large upsampling ratios required for this task (see Section 4.4.1). Using the same underlying BraTS data as in Section 4.3 ensures consistency and avoids distribution shifts when comparing generative and detection models. The availability of ground-truth tumour segmentations further enables quantitative evaluation of segmentation performance, as well as indirect assessment of SR efficacy by analysing lesion-level performance across different resolutions.

In addition to real-world BraTS dataset, the SD dataset generated in Section 4.3.5 is utilised for evaluation purposes to obtain the relationship between lesion size and detection performance. This SD dataset is only utilised for evaluation purposes: all SR and segmentation models have not been trained on any synthetic data.

Degradation pipeline for MRI

As outlined in Section 2.3.5, the degradation pipeline used to synthetically generate paired high-resolution (HR)-LR samples plays a critical role in training SR models, particularly in the absence of naturally paired data. A widely adopted baseline is the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) pipeline, which applies sequential blur, noise, and compression operations to simulate LR image formation. While this approach has been adapted for MRI as shown in Algorithm 1, further refinements may be necessary to capture the specific degradation characteristics of MRI. In this work, the pipeline is extended by modifying the employed noise types and incorporating bias field alterations, thereby more accurately reflecting the degradation processes inherent to MRI data. The refined pipeline is illustrated in Algorithm 2, with parameters detailed in Table B.5, and its impact evaluated in Section 4.4.2.

Algorithm 2 Extended degradation pipeline for MRI by adapting Algorithm 1

Require: High-resolution image tensor $x \in \mathbb{R}^{H \times W}$ with C MRI sequences
Require: Number of degradation levels N
Ensure: Degraded image tensor $\tilde{x}$ at original resolution

$\tilde{x} \leftarrow x$
for $i = 1$ to N **do**
    $\tilde{x} \leftarrow \text{Blur}(\tilde{x}, k_i)$ ▷ Blur Kernel $k_i \sim \{\text{iso, aniso}\}$
    $\tilde{x} \leftarrow \text{Resize}(\tilde{x}, \text{scale} = s_i)$ ▷ Resize scale $s_i \in [0.3, 1.5]$
    $\tilde{x} \leftarrow \text{Noise}(\tilde{x}, \eta_i)$ ▷ Noise parameters $\eta_i \in \{\text{Gaussian, Rician, Non-chi}\}$
    $\tilde{x} \leftarrow \text{BiasField}(\tilde{x}, b_i)$ ▷ Bias field parameters b_i
end for
$\tilde{x} \leftarrow \text{Resize}(\tilde{x}, \text{target size} = H \times W)$ ▷ Resample to original resolution
return $\tilde{x}$

Model architecture and training

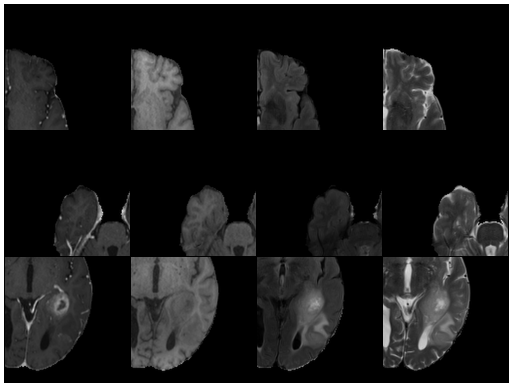
As outlined in Section 4.1 and reiterated throughout this chapter, both small lesion generation and detection using SR share a common architectural backbone, differentiated mainly by their conditioning signals. In contrast to the generative setting, which uses the binary lesion mask as conditioning, the SR task leverages a synthetically degraded LR image as input. The model architecture therefore remains identical to that in Section 4.3, with details in Sections 4.3.2 and 4.3.3. Using the same architecture across tasks ensures a consistent setup and unifies both into a single framework, enabling direct comparison of conditioning mechanisms. This not only streamlines training but also supports comprehensive evaluation of performance in both small lesion generation and SR. The only difference lies in the spatial context provided to the SR model, described below.

Patch-based training and inference A core aspect of the SR pipeline is that state-of-the-art degradation strategies first upsample the degraded image back to the original HR dimensions before applying the learnable SR model (see Section 2.3.5 and Algorithm 1). Consequently, the SR model operates on an already interpolated image and functions primarily as a refinement stage, enhancing fine details rather than performing a global resolution increase.

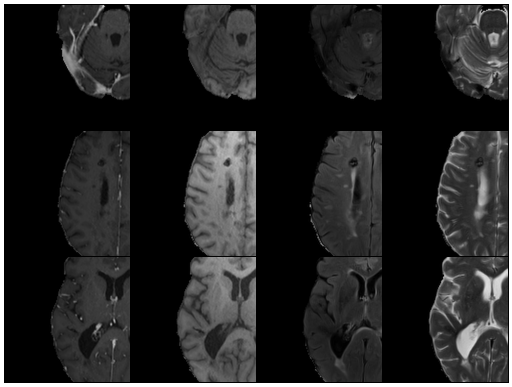
While this approach simplifies architectural integration, it substantially increases the computational burden during inference of the SR model. Attention mechanisms, in particular, become costly as their memory and time complexity scale quadratically with spatial resolution. Additionally, gradient-based anomaly detection techniques (see Section 4.4.1) are sensitive to image size, further compounding the computational demands. To address these constraints, a patch-based inference strategy is adopted. HR images are divided into smaller, overlapping patches that are processed independently, reducing memory requirements at the cost of increased inference time. However, this necessitates training models on patch-based inputs to maintain consistency with the inference procedure. Otherwise, a mismatch arises between training and inference, leading to degraded performance (see Fig. B.7). For this purpose, the PD dataset is employed, which provides 256×256 patches extracted from the original BraTS images (see Section 4.2). The patch size is chosen to balance computational feasibility with sufficient anatomical context for effective SR and lesion detection.

First-stage model re-evaluation The patch-based inference strategy outlined in the previous section necessitates a re-evaluation of the first-stage models, specifically the VQAE and KLAE for image encoding. Both models are retrained on the PD dataset, aligning with the patch-based strategy employed during inference (see Section 4.4.1). As the spatial resolution of the patches remains unchanged relative to the generative setting using full-scale images, the architectural configuration of the first-stage models is preserved. Consequently, the evaluation protocol mirrors that presented in Section 4.3.3, with the sole modification being the use of cropped inputs in place of full-resolution images.

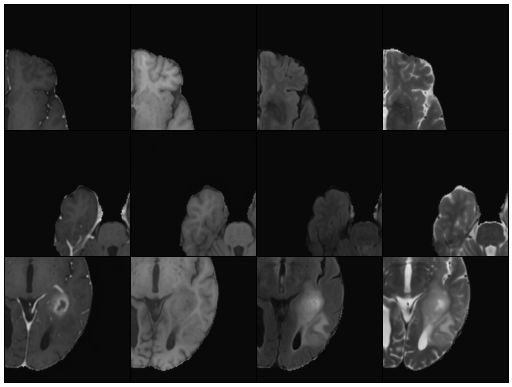
The evaluation of the first-stage models under the patch-based configuration, presented in Fig. 4.11, reaffirms the superior reconstruction quality of the KLAE compared to the VQAE. The KLAE achieves a lower average reconstruction error (0.3323 vs. 0.5486) and preserves visual fidelity more effectively, successfully compressing image crops while maintaining fine anatomical details. In contrast, the



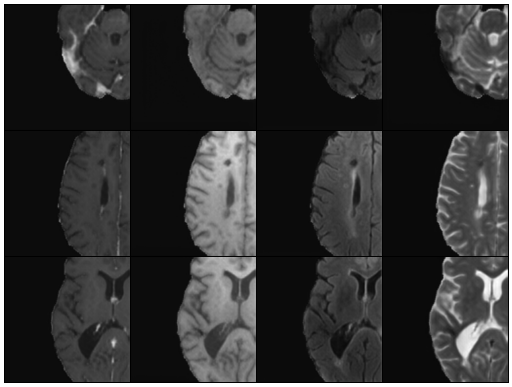
(a) KLAE: Input



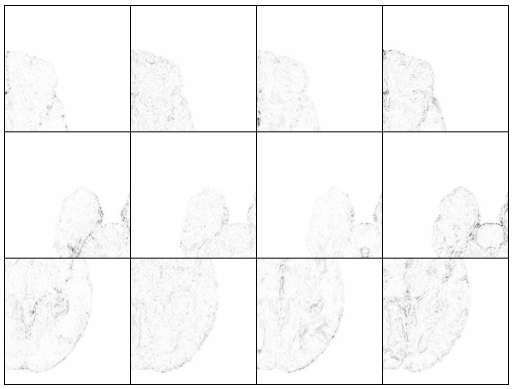
(b) VQAE: Input



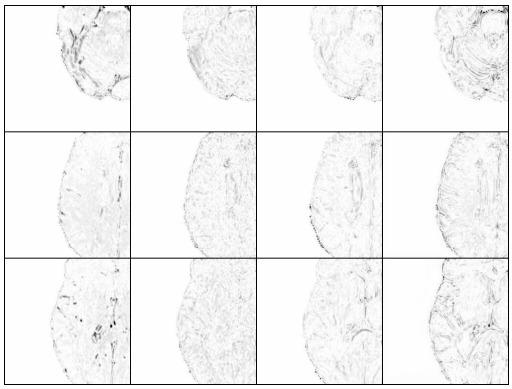
(c) KLAE: Reconstruction



(d) VQAE: Reconstruction



(e) KLAE: Difference



(f) VQ-GAN: Difference

Figure 4.11: Comparison of first-stage models for patch-based SR-LDM. Left column shows results from the KLAE, right column from the VQAE. Rows represent the input images, reconstructions, and absolute reconstruction errors, respectively. Within each subplot, columns correspond to the MRI sequences (ordered as T_1ce, T_1w, T_2-FLAIR, and T_2w), while rows correspond to different 2D slices from individual subjects.

VQAE continues to be limited in high-frequency domains, particularly at lesion and brain boundaries, leading to higher reconstruction errors and visibly reduced clarity. These findings confirm that the KLAE remains the preferred choice for encoding image data into a compact latent space, even in a patch-based setting. It is therefore selected as the first-stage model for subsequent experiments within the LDM framework.

Anomaly map generation

The proposed approach is structured as a sequential pipeline, where the SR module serves as a pre-processing step to enhance spatial resolution and fidelity. This design allows the same weakly-supervised anomaly detection model to be applied consistently across different resolutions, enabling direct comparison of performance. In addition, decoupling the two tasks improves computational efficiency, as the SR and anomaly detection modules can be executed with different batch sizes, reducing inference time for the overall pipeline.

Anomaly map generation detailed in Section 2.3.6 is central to the detection process, yet the commonly used image-space differencing approach is highly sensitive to outliers (see Section 3.6). In small lesion settings, this sensitivity poses a particular challenge, as large differences in healthy regions can dominate the dynamic thresholding and suppress true lesion signals. This limitation highlights the need to explore alternative strategies beyond image-space thresholding.

To approach this problem, I propose a novel gradient-based anomaly map generation strategy called ***GradDiff***. Particularly, two components are combined in an attempt to mitigate outliers and enhance the robustness of the anomaly map generation process:

1. Utilisation of the difference in the latent space between the class-conditional prediction and the unconditional prediction formulating a pseudo-gradient (see Section 2.3.6 and Eq. (2.45)), and
2. the computation of a true gradient in latent space between the predicted output and the encoded representation of the input image.

The combination of both gradients is hypothesised to highlight regions where the model applies substantial modifications to the input, approximating areas of altered probability density. As the gradients are defined in the latent space of the 2D-LDM, they are interpolated to the original input resolution and used to generate a region-of-interest mask. This mask constrains subsequent intensity-based thresholding to lesion-relevant areas, reducing the influence of outliers in non-lesion regions.

Empirical analysis showed that gradient magnitudes decrease as the sampling trajectory progresses, with the most informative updates occurring early (see Figure B.8). Accordingly, gradients were tracked during the initial timesteps of the sampling trajectory to maximise information. A similar pattern was observed for class-conditional versus unconditional differences, motivating the same strategy. The final hyperparameters for gradient computation were determined through a systematic grid search, with full details given in Table B.6.

Given the large configuration space (≈2.5 million permutations), only the principal outcomes of the grid-search are reported here. Optimal performance was achieved with $N = 600$ sampling steps and a classifier-free guidance scale of $C = 5$. From the recorded timesteps, aggregating five evenly spaced points from the final 10% of the trajectory using the median produced the most robust anomaly maps. Mean aggregation across MRI channels was consistently effective, and image-space morphological post-processing (opening followed by dilation) further improved mask coherence. In contrast, latent-space morphological operations provided no consistent benefit.

Evaluation of small lesion detection

The evaluation investigates the effect of SR on lesion detection performance using a weakly-supervised 2D-LDM trained for anomaly detection. In the detection pipeline, this model succeeds the SR LDM, as described in Section 4.4.1. It builds on the findings of the previous chapter (see Chapter 3) and aligns with state-of-the-art weakly-supervised methods (see Section 2.3.6). Full configuration details are provided in Table B.7.

Several baselines are included for comparison. The weakly-supervised models from the previous chapter, namely 2D-variational autoencoder (VAE), 2D-class activation map (CAM) (Z. Chen et al., 2022), and 2D-WS-MTST (H. Chen et al., 2023), are adapted to the lower-dimensional setting. Finally, a supervised U-Net trained with a combined cross-entropy (CE) and Dice loss serves as a reference, providing an upper bound on performance under the same preprocessing conditions.

The evaluation setup relies on the dataset variants introduced in Section 4.2. All detection models, including baselines, are trained exclusively on the same FD dataset. This design enables direct assessment of lesion detectability at both standard and super-resolved resolutions using the same 2D-LDM, ensuring comparability between the two settings. Evaluation is performed on the super-resolved images with interpolated ground-truth masks for the SR approaches, as preliminary experiments showed this to be equivalent to downscaling them back to the original resolution. For all non-SR models, evaluation is carried out directly at the original resolution. Detection performance is independently assessed on two datasets: (1) the held-out test set from the FD dataset, containing real lesions from the BraTS dataset, and (2) the SD dataset (see Section 4.3.5), which contains controlled small lesions of diameters $d_d \in 5, 10$mm. For evaluation, 20 subjects with five 2D slices are randomly selected from each dataset. The test set size is limited by the computational cost of the SR and segmentation pipelines (see Section 4.4.4). To address this, performance is estimated through bootstrapping, drawing 10 resampled sets from 100 available samples. This procedure enables statistical assessment and provides more reliable generalisation estimates.

Evaluation metrics include the Dice similarity coefficient (DSC) and specificity, chosen for their clinical relevance. DSC measures the spatial overlap between predictions and ground truth, providing a standardised assessment of delineation accuracy that remains sensitive to the alignment of small volumes. Specificity complements this by measuring the ability to avoid false-positive detections. This is particularly important in the clinical setting of small lesion detection, where over-segmentation may lead to unnecessary concern or intervention (see Section 2.1.2).

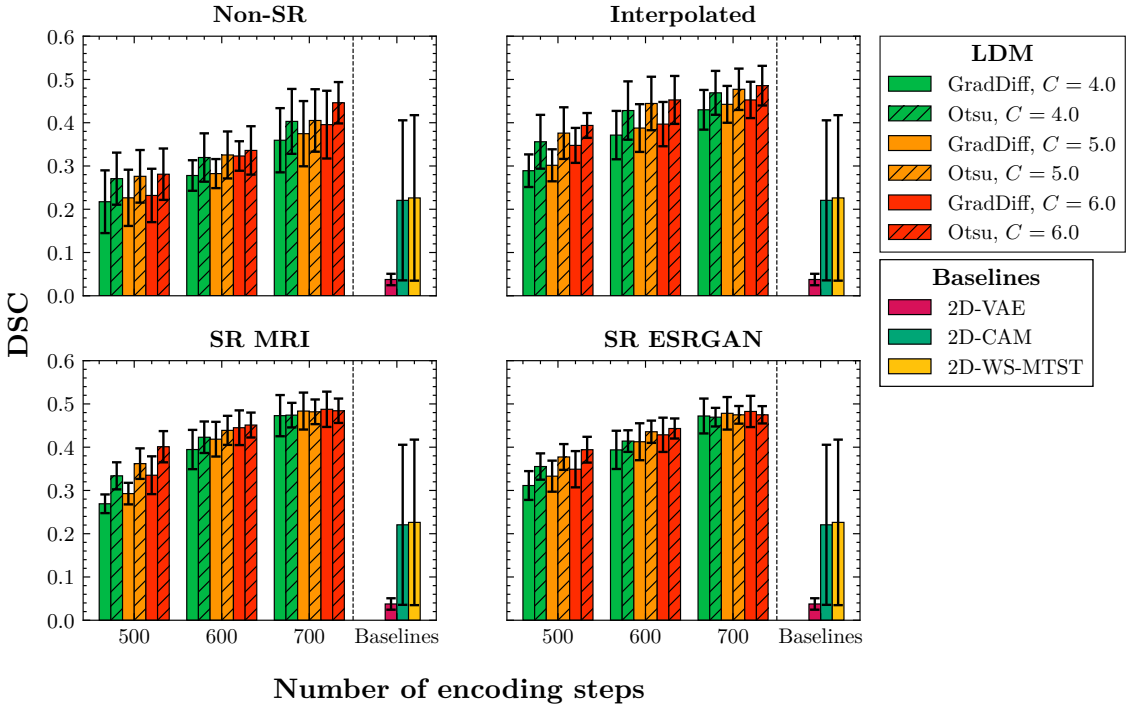


Figure 4.12: Performance for the FD dataset. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

4.4.2 Lesion detection experiments

The following section presents the results of the SR experiments, focusing on the impact of spatial resolution on small lesion detection. The section first investigates the effect of SR on detection performance across different lesion sizes, followed by an ablation study investigating the thresholding mechanism and the detectability of non-synthetic small lesions.

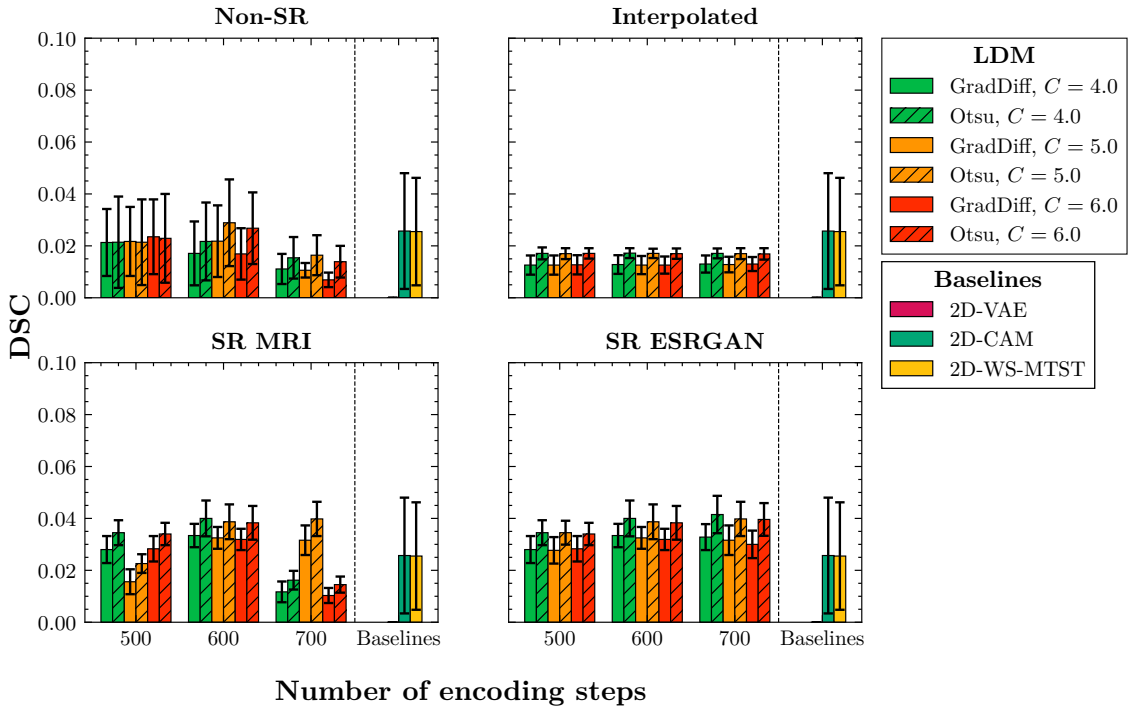


Figure 4.13: Performance for the SD dataset with 10 mm lesion diameter. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

Important notes

Baseline methods are evaluated only on the non-SR FD dataset. The same results are shown across all plots to facilitate direct comparison with the 2D-LDM operating at different resolutions.

Baseline results

Real lesions of various sizes (FD) The 2D-VAE performs poorly in detecting real lesions of various sizes, with a very low DSC of 0.0376 ± 0.0132 (see Fig. 4.12). Specificity is high at 0.9983 ± 0.0007 as shown in Fig. B.13, which reflects a strong bias towards background predictions. The obtained scores clearly indicate the lack

of label guidance during training, leading to inferior detection performance.

2D-CAM outperforms the VAE in detection, achieving a DSC of 0.2206 ± 0.0185 and moderate specificity of 0.6464 ± 0.4817. The transformer-based 2D-WS-MTST performs slightly better, with a DSC of 0.2262 ± 0.1913 and specificity of 0.6522 ± 0.4811. Both models, however, show high uncertainty, indicated by wide confidence intervals, suggesting unstable performance across samples.

In contrast, the 2D-LDM without SR surpasses all other weakly-supervised baselines in the evaluated settings. The best configuration is obtained with traditional thresholding, $N = 700$ and $C = 6$, achieving a DSC of 0.4462 ± 0.0478 with very high specificity 0.9994 ± 0.0004. All tested hyperparameter configurations are characterised by elevated performance compared to other baselines, reduced variance across samples and high specificity (see Figs. B.13 and 4.12). DSC increases with higher N and C, showing better average predictions at the cost of less stable performance indicated by growing variances. The GradDiff thresholding method, designed to suppress outlier responses, consistently underperforms Otsu thresholding, with a reduction in DSC of about 4 % to 5 % across all configurations. This performance gap narrows with increasing N and C.

Synthetic lesions (SD) For synthetic lesions of 10 mm, weakly-supervised baselines without diffusion remain limited yet show modest detection capability. 2D-CAM and 2D-WS-MTST achieve comparable DSC (0.0257 ± 0.0223 and 0.0255 ± 0.0207, respectively) with moderate specificity (0.6233 ± 0.4866 vs 0.6584 ± 0.4793). In comparison, the 2D-VAE fails to segment synthetic targets altogether, with DSC 0.0000 and specificity 0.9961 ± 0.0025 reflecting near-uniform background predictions.

For synthetic lesions, the advantage of the 2D-LDM without SR over the other baselines is less pronounced than for real lesions, with the exception of the 2D-VAE. Performance decreases across all tested configurations (see Fig. 4.13) and is accompanied by relatively high variance across samples. The best DSC for 10 mm lesions of 0.0289 ± 0.0167 is achieved with $N = 600$ and $C = 5.0$ using Otsu thresholding. Larger N no longer improves results, and increasing C yields inconsistent effects.

Furthermore, GradDiff performs comparably to Otsu thresholding. The obtained specificities remain at almost ideal values (see Fig. B.14), indicating a strong bias towards background predictions, which is the likely cause of the low DSC scores.

Further reducing the lesion size to 5 mm leads to a complete failure of all weakly-supervised baselines, including the 2D-LDM without SR. All models yield DSC scores of 0.0000, with specificity values close to 1, indicating uniform background predictions (see Figs. B.9 and B.15). This outcome highlights the inherent limitations of weakly-supervised methods in detecting very small lesions, likely due to insufficient signal strength and the absence of explicit lesion supervision during training.

Importance of higher resolution

Real lesions of various sizes (FD) Increasing the effective resolution of the 2D-LDM through conventional interpolation yields substantial improvements compared to all baselines across all tested configurations (see Fig. 4.12, "Interpolated"). On average, the DSC rises by 0.0842 for GradDiff and 0.0975 for Otsu thresholding relative to the non-SR baseline. The best results are achieved at $N = 700$ and $C = 5$, with DSC scores of 0.4425 ± 0.0426 for GradDiff and 0.4773 ± 0.0478 for Otsu. Variances are reduced compared to the baseline, indicating more stable and confident predictions. Specificity remains consistently high across all configurations (see Fig. B.13), confirming a low false-positive rate.

Applying SR refinement to the interpolated images provides additional but smaller gains as shown in Fig. 4.12, bottom row. Using the MRI degradation pipeline, the average DSC improves by 0.0225 for GradDiff, while Otsu thresholding shows negligible change. The best configuration is obtained with GradDiff at $N = 700$ and $C = 6$, reaching a DSC of 0.4876 ± 0.0408. Variances are further reduced compared to the interpolated samples, reflecting improved robustness to sample variation and outliers. The choice of the SR degradation pipeline has minimal impact on performance, as outlined in Section 4.4.2.

As with the baseline, higher N and C generally improve performance, with larger N yielding more stable results. The gap between GradDiff and Otsu thresholding

narrows under these settings, with GradDiff achieving the best overall score at the expense of slightly higher variance in most configurations. Performance also appears to plateau, as further increases in N and C produce diminishing improvements compared to the non-SR baseline.

In summary, increasing spatial resolution through interpolation markedly enhances the detection performance of real lesions with varying sizes obtaining substantially better results than all weakly-supervised baselines and substantial improvements for the 2D-LDM. Further refinement using SR provides additional but smaller gains, suggesting that while SR can enhance image quality, the primary benefit arises from the initial resolution increase. The results also highlight the effectiveness of Otsu thresholding in this context, with GradDiff offering marginal improvements primarily at higher N and C settings.

Synthetic lesions (SD) Performance gains of SR over conventional interpolation are also observed for synthetic small lesions, with comparable improvements in DSC (see Fig. 4.13). However, variances are elevated relative to the interpolated counterpart, complicating reliable performance estimates. Moreover, the trend of higher N and C improving results is absent, with performance stagnating across the tested range. This effect becomes more pronounced for smaller synthetic lesions (see Fig. B.9). While SR still outperforms the baselines, regular interpolation and non-SR 2D-LDMs in terms of both DSC and variance reduction, the results highlight the approaches insufficient ability to consistently detect very small lesions in the tested setting.

Influence of degradation pipeline

Comparing the two degradation pipelines within the context of SR, no substantial differences are observed in detection performance across any of the tested configurations (see Figs. B.13 and 4.12). The MRI-specific pipeline exhibits slightly reduced variance in DSC scores when conventional thresholding is used, suggesting a marginal benefit in stability. However, this effect is not replicated when using GradDiff,

which appears unaffected by the choice of degradation pipeline, particularly in real lesion settings. Specificity remains effectively constant across all configurations and approaches, reflecting the already strong bias of the models towards background prediction. These results indicate that, while additional MRI-specific artefacts were introduced to better approximate clinical degradation, they do not meaningfully alter the model's capacity to localise lesions in the evaluated SR scenario.

For synthetically generated small lesions with a diameter of 10 mm, the comparison between degradation pipelines shows slightly more pronounced differences (Fig. 4.13, bottom row). The ESRGAN-based pipeline exhibits greater stability, with less fluctuation across different settings of C and N, and a clearer trend of improved performance as N increases. In contrast, the MRI-specific pipeline shows more variability. However, the absolute differences in DSC remain very small and likely fall within the normal variability of the evaluation. Thus, despite the apparent trends in the plot, the overall conclusion is consistent with the real lesion analysis: both degradation strategies yield comparable detection performance in the SR setting.

An analysis of 5 mm synthetic lesions was omitted in the interest of time. Previous experiments with both real and synthetic data demonstrated consistent performance across degradation pipelines, indicating that additional evaluation at this scale would be unlikely to provide further insight.

Summary Overall, the results between pipelines remain within expected variance margins, indicating that the adapted ESRGAN pipeline is a suitable choice for the BraTS dataset. While no measurable advantage was observed, the inclusion of MRI-specific artefacts, including bias field inhomogeneities and domain-relevant noise, positions the novel pipeline introduced in this thesis as a viable candidate for future application to real-world clinical data. The preservation of performance across both pipelines further supports its suitability as a step towards more generalisable and robust refinement of clinical images.

Comparison against supervised U-Net

To contextualise the performance of the presented weakly-supervised approaches, a fully supervised segmentation model was included as an upper-bound reference. This decision was motivated by the consistently underwhelming performance observed across all tested methods on the synthetic small lesions. In addition to highlighting this performance gap, the comparison serves to quantify the trade-offs between supervision level and segmentation quality. It is important to note that the supervised model operates under fundamentally different training conditions, relying on fine-grained voxel-level annotations for direct guidance. While these labels offer strong supervision, their acquisition is labour-intensive, subject to inter-rater variability, and often infeasible at scale, as discussed in Section 2.1.8.

As shown in Figs. B.10 to B.15, the supervised U-Net markedly outperforms all weakly-supervised methods across all lesion sizes. For real lesions, the U-Net almost obtains a perfect DSC with 0.9580 ± 0.0058 and a specificity of 0.9989 ± 0.0001. This performance starkly contrasts with the best weakly-supervised LDM configuration, which achieves a DSC of 0.4876 ± 0.0408 and specificity of 0.9994 ± 0.0004 using SR.

The performance gap remains evident for synthetic lesions, though less pronounced than in real data. For 10 mm and 5 mm synthetic lesions, the supervised U-Net achieves DSC scores of 0.2904 ± 0.0292 and 0.1330 ± 0.0194, respectively. These results confirm that the synthetic lesions encode features that are at least partially detectable under strong supervision. However, the segmentation quality remains markedly lower than for large real lesions, where the same model approaches near-perfect performance.

This discrepancy may point to inherent limitations in the generative process: synthetic lesions may occasionally fail to manifest a clear lesion in the image due to guidance signals vanishing during compression, poor integration into the latent representation, or insufficient fidelity in conditioning. These issues are not visually apparent and cannot be exhaustively verified by me due to limited experience in the manifestation of real small lesions in MRI. The unexpectedly low performance of the

U-Net in this setting, however, motivates a targeted ablation study on real small lesions of the BraTS dataset to determine whether these shortcomings originate from the generative model or reflect intrinsic challenges associated with subtle lesion characteristics.

4.4.3 Ablation study: non-synthetic small lesions and thresholding investigation

To investigate the limitations suggested by the reduced U-Net performance on synthetic lesions, a targeted ablation study was conducted. The primary aim is to evaluate whether the generative model occasionally produces synthetic samples in which no lesion is meaningfully embedded (see Section 4.4.2). In parallel, the study assesses the impact of the thresholding mechanism. Specifically, the GradDiff method showed promise in SR scenarios, but its performance was inconsistent across configurations and has the tendency to underperform compared to Otsu thresholding (see Section 4.4.2). This motivates further exploration of whether its sensitivity can be refined to better capture small lesions and whether these benefits can be recovered without disproportionate computational overhead (see Section 4.4.4). Combining both aspects within one ablation study allows for a comprehensive evaluation of the model's capabilities and limitations in detecting small lesions in a controlled setting.

The analysis is conducted on a subset of the FD dataset. Instead of randomly sampling slices from the test split, this ablation selects samples containing lesions with a diameter of 10 mm. As outlined in Section 4.3, such cases are rare due to the variability of lesion sizes in the dataset. Consequently, only one slice per subject from 20 test subjects meets these criteria. Although this limits statistical power, it enables controlled evaluation on real data, which is essential in this context. This "new" dataset is re-evaluated using the same methodology as in Section 4.4.1.

The ablation study begins with the supervised 2D U-Net, which provides an upper-bound reference for comparison. On the real dataset, the U-Net achieves a Dice of 0.4765, whereas performance drops markedly to 0.2904 on the synthetic dataset. This suggests that real lesions are generally easier for the pre-trained model

to detect. At the same time, the results support the hypothesis that certain synthetic lesions may be absent or difficult to represent, although this cannot be fully verified without expert annotation. It remains equally possible that the lesions exist but are harder to delineate, highlighting the ambiguity of borderline cases.

The 2D-LDM with SR was subsequently evaluated on the real dataset using the optimal GradDiff parameters identified in the preceding experiments (see Section 4.4.1), using $N = 500$, $C = 4$. The Dice score of 0.0378 closely matches the result obtained on the synthetic dataset (0.0334), with the small difference falling within the expected variability of the evaluation. This indicates that the segmentation performance of GradDiff remains stable when transitioning from synthetic to real lesions.

To investigate whether more tailored hyperparameter choices could enhance performance, the same grid search as outlined in Section 4.4.1 was conducted to explicitly optimise GradDiff for the given dataset. A substantial gain was observed for GradDiff at $N = 500$, $C = 4$, where the Dice increased to 0.0960, surpassing Otsu under the same conditions, and doubling the DSC. These results were obtained by relying exclusively on the gradient of the healthy prediction, with the choice of aggregation method and number of timesteps having little effect. In contrast, if the memory-intensive latent gradient computation is entirely disabled and only the pseudo-gradient from classifier-free guidance is used, performance is only slightly reduced, reaching a DSC of 0.0837. These results indicate that GradDiff can be more effective for small lesions when carefully tuned.

In summary, the ablation study demonstrates two key findings: (1) supervised baselines perform better on real lesions than on synthetic ones, which may indicate that certain synthetic lesions are harder to represent or delineate reliably, and (2) careful hyperparameter optimisation recovers lost sensitivity of GradDiff for small lesion detection with SR, surpassing established thresholding methods and doubling performance for 10 mm lesions.

Table 4.5: Runtime for small lesion detection across different input types. Inference time is reported in seconds per slice of the test dataset, separated into segmentation and SR. N denotes the number of encoding steps. The interpolation time for "Upscaled" is negligible.

Input type	SR (s)	N	Segmentation (s)	Total (s)
Non-SR	0.0	500.0	110.0	110.0
		600.0	131.77	131.77
		700.0	152.67	152.67
Interpolated	0.0	500.0	1391.08	1391.08
		600.0	1628.9	1628.9
		700.0	1859.49	1859.49
SR	1654.56	500.0	1371.85	3026.41
		600.0	1653.67	3308.22
		700.0	1886.07	3540.62

4.4.4 A word on inference times

Despite promising results across configurations, SR and upscaled images exhibit much longer inference times than the non-SR baseline, as shown in Table 4.5. The 4× resolution increase causes a considerable rise in computational cost, both from the exponential growth in predictions and the added expense of generating gradient-based anomaly maps with GradDiff. Inference on upscaled images was only feasible with a batch size of 1, even on a high-end Nvidia A40 with 40 GB memory. This underscores a key limitation in scalability and raises questions about whether the performance gains justify the heavy computational burden.

Reported timings in Table 4.5 should be considered approximate, as measurements were obtained across different server configurations and include data loading times. The latter is strongly influenced by CPU load and access delays due to shared file system infrastructure across all nodes on the computer cluster. Nonetheless, the overall trend remains clear: combining segmentation with SR more than doubles the inference time relative to standard upscaling. The full-resolution SR configuration results in total inference times approximately 30 times higher than the baseline, reaching roughly 50 min per slice. This level of resource demand was only manageable

by decoupling the two stages: SR outputs were written to disk and processed separately during segmentation. This separation allowed larger SR inference batch sizes due to the absence of GradDiff. These resource constraints explain the limited evaluation depth: with inference times of approximately 50 min per slice, testing additional parameters or larger datasets was computationally infeasible.

Based on the results of the ablation study in Section 4.4.3, one option to reduce memory usage is to disable latent gradient computation in GradDiff and instead rely on the pseudo-gradient from classifier-free guidance. This configuration substantially decreases memory requirements, allowing larger segmentation batch sizes and thereby reducing inference time, while only marginally affecting performance on the tested dataset. A broader assessment of its behaviour on other data lies beyond the scope of this thesis and remains a subject for future investigation.

Patch-based prediction was also explored as a complementary strategy to reduce resource demands and enable inference on the available hardware. While it allowed larger batch sizes and lower memory consumption, segmentation performance dropped substantially, likely due to the reduced spatial context within each patch and the challenges of merging overlapping tiles. Training a dedicated patch-based model on fully upscaled data may mitigate these issues, but this was outside the scope of the current work. Nonetheless, both strategies remain promising directions for improving the efficiency and accessibility of high-resolution inference.

4.4.5 SR and anomaly detection unification

In light of the increased inference times discussed in Section 4.4.4, unifying SR and anomaly detection into a single model presents an attractive direction to improve efficiency. Such a framework would allow both HR reconstruction and healthy counterfactual generation within a single forward pass, eliminating sequential processing. A theoretically feasible approach involves incorporating class-conditioning into the structural encoder of the pre-trained weakly-supervised LDM. Specifically, the structural encoder would be extended with label embeddings and conditioning mechanisms, as described in Section 2.3.3. This would enable both the SR encoder

and the segmentation backbone to operate under class-conditional guidance consistent with the classifier-free formulation outlined in Section 2.3.6. During inference, the model would be tasked with both resolution enhancement (structural encoder) and conditional healthy counterfactual generation of the input (backbone LDM).

However, this integration poses a fundamental challenge as shown in Fig. B.19. Because the SR module precedes the anomaly detection backbone, it always operates on the LR input in which the lesion may be present. This persistent exposure to pathological content at each sampling step overrides the subtle conditional guidance required for healthy counterfactual generation. The conflicting objectives of increasing resolution of a diseased input while simultaneously conditionally removing the lesion components proved difficult to reconcile. In practice, the injected class-conditioning signal was insufficient to counteract the dominant pathological features in the input, resulting in SR outputs with minimal or no visible alteration. Consequently, while the unified architecture is conceptually appealing, its current implementation is unable to effectively disentangle resolution enhancement from semantic modification.

4.4.6 Summary of key findings

Increased resolution improves segmentation accuracy as indicated by DSC, while also reducing variance and false positive predictions. This demonstrates enhanced robustness when applying the same anomaly detection model trained on MRI at original resolution. Conventional upsampling alone provides a performance gain of almost 10 % across all LDM configurations. Refining the upscaled image with a dedicated SR module yields a further 2 % improvement, substantially surpassing conventional weakly-supervised baselines. The degradation pipeline plays a critical role in enabling these refinements, as it allows learning of the inverse of the synthetic degradation process. However, it must be adapted to the degradations present in the target data. State-of-the-art pipelines already integrate such steps, and further refinements were not explicitly required for the BraTS dataset. Nonetheless, data quality remains paramount since SR models are highly sensitive to distribution shifts, as initially observed on synthetic data.

Detecting small lesions remains a critical challenge. SR offers improvements but is constrained by the anomaly detection mechanism, which relies on intensity-based differences and is therefore vulnerable to outliers. The gradient-based thresholding mechanism GradDiff was introduced to address this limitation. When carefully tuned, GradDiff delivers notable performance gains for small lesions. These improvements, however, depend on specific parameter choices and their generalisability remains uncertain. The evaluation on both real and synthetic small lesions underscored the difficulty of reliably detecting such subtle abnormalities across all approaches, including supervised methods. The results demonstrate that these limitations are not solely due to the level of supervision, but reflect a broader challenge that requires alternative strategies to achieve consistent detection of small yet clinically important lesions.

4.5 Limitations

Despite the promising results achieved in this work, several limitations and challenges remain that warrant further discussion.

The first limitation of this study lies in the use of downscaled large lesions to simulate small lesions. While this strategy preserves structural plausibility, it may not accurately capture the true morphology or intensity characteristics of naturally occurring small lesions. The visual appearance, boundary definition, and contextual integration of true small lesions may differ substantially. Furthermore, the rounding required during spatial scaling introduces discretisation artefacts, particularly for very small lesions.

A further consideration arises from the possibility that very small or faint lesions may not be preserved during the downsampling process applied by the first-stage model of the generative LDM. As a result, such lesions may be partially or entirely removed in the latent space and subsequently will not appear in the generated imaging. This may render them generally undetectable by the segmentation model, as they are not present in the input data. Consequently, it is difficult to determine

whether the model failed to detect the lesion or whether the lesion was imperceptible in the latent representation to begin with. As a non-clinician, a definitive evaluation of lesion visibility in these borderline cases is not feasible (see Section 4.3.5) and beyond the scope of this work.

A primary focus of the performance evaluation was the detection of lesions at the 10 mm threshold, aligning with the RECIST criteria for measurable disease (see Section 2.1.6). While this provided a rigorous stress test for the framework at the extreme lower limit of clinical relevance, it also highlighted the inherent sensitivity of overlap metrics to minor spatial discrepancies in small volumes. Evaluations conducted at even smaller scales, such as 5 mm, did not yield further meaningful evidence, as the 10 mm boundary already represents the current sensitivity threshold of the approach. Future research could further delineate the model's performance envelope by evaluating a broader range of larger lesions to determine the point at which detection sensitivity plateaus. This was not feasible in the current study due to the substantial computational requirements and inference times necessitated by the high-dimensional hyperparameter search and multi-dataset evaluation.

Despite their flexibility, DDPMs remain computationally expensive, especially in their standard formulation. More efficient variants have been proposed (H. Chung et al., 2022, 2023), but were not explored here to allow for a more consistent comparison to established methods. In addition, efficient sampling usually comes at the cost of reduced performance, which can be crucial in the context of detecting smaller distribution shifts robustly. Moreover, large parameter spaces, particularly for conditioning and sampling configurations, complicate systematic experimentation. Long training times make exhaustive testing impractical, and minor configuration changes introduced during development may affect results.

Furthermore, the 2D-LDM used for anomaly detection was trained on the original resolution data but applied to SR samples to maintain comparability across experiments. This ensured consistent evaluation with non-SR data but may not be optimal for higher-resolution inputs, as the model was not designed to capture the finer detail introduced by SR. Consequently, faint or small lesions could still be

missed. Similarly, the first-stage models were only trained on HR data, yet were applied to encode LR samples during the training of the SR module. Adapting these models to the specific resolution levels used in SR could improve reconstruction fidelity and downstream segmentation performance, at the cost of largely increased training complexity and time.

The attempted unification of SR and segmentation within a single class-conditional generative model to mitigate the computational constraints revealed a critical limitation: the SR module, positioned prior to the segmentation backbone, receives the low-resolution, lesion-containing image as input at every diffusion step. This strong pathological signal dominates the conditioning pathway, resulting in outputs with minimal or no visible alteration. The injected class information proved insufficient to counteract the lesion-preserving guidance of the SR model, ultimately preventing effective counterfactual generation.

The proposed GradDiff approach for anomaly map generation was designed to improve robustness and accuracy in the segmentation process. Its reliance on latent gradients aims to minimise outliers that commonly affect intensity-based thresholding. However, computing real gradients in latent space is computationally expensive and demands considerable memory resources, especially for high-resolution images where latent representations become large. The ablation study in Section 4.4.3 shows that careful hyperparameter tuning can substantially improve segmentation performance for small lesions. Moreover, it demonstrates that the real gradient may not be strictly required, as relying solely on the pseudo-gradient from classifier-free guidance leads to only a marginal reduction in performance for GradDiff. This suggests that GradDiff retains its effectiveness even under memory-efficient configurations, making it a promising approach for practical applications. Nonetheless, the introduction of additional parameters increases model complexity, and their general benefit across datasets remains uncertain.

MRI remains the gold standard for brain tumour assessment due to its high soft-tissue contrast and spatial resolution. However, image quality and lesion visibility are strongly affected by acquisition parameters, scanner characteristics, and artefacts

(Eisenhauer et al., 2009). These factors introduce additional complexity when evaluating detection models, particularly for small or subtle lesions. Utilising a larger multi-institutional dataset with varying morphologies could be useful to accurately adapt the flexible degradation pipeline presented in this work. This would allow to compensate for specific artefacts prior to the segmentation model, and thereby improve the robustness of the segmentation model, as demonstrated in this work.

Lastly, while image resolution and segmentation accuracy are essential, they represent only one aspect of early tumour detection. Broader diagnostic timelines are also influenced by healthcare access, clinical suspicion, and neuroimaging availability (Dobrovoljac et al., 2002). The work presented here addresses the technical feasibility of lesion detection under constrained supervision but should be interpreted as one part of a broader clinical process.

4.6 Conclusion

DDPMs have demonstrated considerable versatility throughout this chapter. They were applied successfully to synthetic data generation (Section 4.3), SR (Section 4.4), and anomaly detection (Section 4.4.2) all by modifying the conditioning signal. This shared foundation enables the construction of unified frameworks capable of addressing diverse objectives with minimal architectural divergence. In this chapter specifically, I have demonstrated the benefits of SR for the detection of small lesions, achieving substantial performance gains over established weakly-supervised baselines. In addition, the evaluation was only possible using a synthetic dataset with precise lesion characteristics. Both in combination address the second research gap of this research project (see Gap 2). Specifically, I:

- Developed a synthetic model for generating brain tumours of predefined lesion size and location, enabling robust and reproducible evaluation of lesion detection performance.
- Investigated SR techniques to improve lesion visibility, enhancing segmentation accuracy, robustness, and sensitivity to smaller lesions.

- Evaluated the SR-LDM framework against state-of-the-art methods, assessing its ability to detect and segment small lesions within the synthetically generated dataset and determined the minimum detectable lesion size.
- Introduced and evaluated GradDiff: a novel gradient-based anomaly map generation method to improve segmentation robustness and accuracy, particularly for small lesions.
- Unified synthetic generation and detection of small brain tumours within a single model framework, using a common conditioning mechanism that integrates both generative and detection tasks.

These efforts culminate in the following contributions:

Contribution 2

Developed and validated a DDPM-based model for the controlled synthesis of size-specific brain tumour lesions in multi-sequence MRI.

Contribution 3

Conducted the first systematic investigation of DDPM-based SR in weakly-supervised anomaly detection, evaluating its impact on the detectability of small brain tumours.

Chapter 5

Generalisability to paediatric populations

After the successful investigation of adult brain tumours in 3D (see Chapter 3) and the detection of small lesions in 2D (see Chapter 4), this chapter focuses on the generalisability of the proposed weakly-supervised anomaly detection method to paediatric populations. The emphasis lies on the reusability of previously trained 2D-LDM and components with minimal adaptation, aiming to evaluate how well the approach transfers to the distinct characteristics of paediatric brain tumour cases. The chapter thereby addresses directly the challenges described in Section 2.4.3.

5.1 Introduction and problem formulation

Paediatric brain tumours present unique challenges due to their distinct characteristics compared to adult cases. These differences include variations in tumour types, growth patterns, and the impact of treatment on developing brains, as shown in Section 1.1. Being able to leverage models trained on the comparatively data-rich adult population offers a promising pathway to overcome the amplified data scarcity in paediatric neuro-oncology. Such transfer not only enables the reuse of robust, well-tested components but also helps bridge the gap created by limited research and annotation resources in paediatric cohorts:

Gap 3

The generalisability of weakly-supervised DDPMs to paediatric brain tumours remains untested, despite their hypothesised robustness to distributional shifts and suitability for data-scarce clinical settings.

The chapter is guided by the objective of assessing how well the previously developed models and components from Chapter 3 and Chapter 4 generalise to the paediatric domain. Rather than developing a new model from scratch, the focus lies in reusing and minimally adapting existing elements to account for population-specific differences. This approach reflects the broader motivation to bridge the gap in paediatric neuro-oncological imaging by leveraging knowledge derived from the more data-rich adult population. This idea is formalised in the following research question:

Research Question 3

To what extent can DDPMs trained on adult brain tumour data generalise to the paediatric domain, and how robust are the learned representations to shifts in population and disease distribution?

Answering this research question has significance beyond its immediate relevance to paediatric neuro-oncology. Demonstrating successful generalisation would provide evidence that the proposed weakly-supervised anomaly detection framework can extend to other data-scarce settings with related structural characteristics. This would suggest that knowledge gained from well-annotated populations may be (partially) transferable to clinically distinct cohorts, thereby reducing the need for extensive retraining. More generally, it opens the door for future work exploring how weakly-supervised DDPMs can be adapted to rare diseases or underrepresented cohorts where annotated data is limited, yet diagnostic accuracy remains critical. These findings are summarised in the following contribution:

Contribution 4

Demonstrated the generalisability of DDPM-based weakly-supervised anomaly detection for brain tumour segmentation to data-scarce paediatric cases, validated on a curated multi-institutional dataset under clinically realistic conditions.

The remainder of this chapter is structured as follows: Section 5.2.1 describes the dataset used for evaluation and the metrics employed to assess model performance. Section 5.2 evaluates the generalisability of the proposed framework to paediatric populations using the recently released BraTS paediatric dataset. In particular, it analysis the performance of a pre-trained model and explores various fine-tuning strategies to adapt the model to the paediatric domain. Section 5.3 outlines the collection of a private dataset for evaluation encompassing a wider range of brain tumour subtypes, and presents qualitative and quantitative results for the pre-trained model. Finally, Section 5.4 summarises the findings and discusses limitations.

5.2 Evaluation on the state-of-the-art paediatric dataset

This section evaluates the generalisability of the proposed weakly-supervised framework to paediatric populations using the recently released BraTS paediatric dataset (see Section 2.1.1).

5.2.1 Dataset and evaluation

As outlined in Section 2.1.1, the recent BraTS challenge includes a paediatric cohort of 99 individuals with ground-truth annotations. Although structurally consistent with the adult dataset in terms of available MRI modalities, it is substantially smaller than the 1251 annotated adult cases used to assess low-grade gliomas (LGGs) and high-grade gliomas (HGGs). The dataset offers a valuable benchmark for evaluating how models trained on adult brain tumour data transfer to paediatric cases. Its design closely mirrors the adult BraTS dataset, employing comparable case collections

and consistent processing guidelines during curation. This alignment enables direct analysis of generalisability using pre-trained adult models, with the primary difference being the underlying population. Training models from scratch on the paediatric subset is hindered by limited sample size and variability, leading to overfitting and restricting the capacity of deep feature extractors (Ayana et al., 2024; Ganatra, 2025). Knowledge transfer from the adult domain therefore provides a more robust, data-efficient strategy (Ayana et al., 2024).

Section 5.2.2 investigates the performance of weakly-supervised anomaly detection based on a 2D-LDM without fine-tuning. All 99 paediatric cases are directly used for inference with the 2D-LDM pre-trained on the adult BraTS dataset. In contrast, Section 5.2.3 fine-tunes the proposed 2D-LDM on the paediatric dataset. Due to the limited sample size and the goal of minimal adaptation, the 99 cases are partitioned into 70 for training, 9 for validation, and 20 for testing.

Given the identical structural composition of the paediatric and adult datasets, the data processing pipeline mirrors that described in the preceding chapters. As justified in Section 4.1.3, the analysis is restricted to 2D slices. Each slice is padded to a base-2 resolution of 256×256 pixels and normalised to the range $[-1, 1]$.

The inference procedure is identical to the one described in Section 3.4.3: Encoding is utilised to generate a latent representation of the input data, which is then subject to classifier-free guidance to produce a healthy counterfactual. The segmentation mask is derived from the difference between the original and counterfactual images and thresholded to isolate the lesion region. The thresholding method is selected based on the findings from Chapter 4, where Otsu's method was shown to be effective for small lesions. However, as an alternative, the gradient-based ***GradDiff*** method introduced in Section 4.4.1 is also considered to analyse the benefits of gradient-based thresholding on distribution shifts. To streamline evaluation, only five slices containing lesions are extracted per subject, resulting in 495 slices for the pretrained evaluation and 100 slices for the fine-tuned configuration.

The model's performance is evaluated using the same metrics as in Chapters 3 and 4, namely the DSC and specificity. These metrics are selected for their clinical

relevance, as they quantify segmentation accuracy and the model's ability to avoid false positives. Both are critical for reliable treatment planning and disease monitoring. Their continued use ensures consistency across experiments and reflects the rationale outlined in Section 2.1.2. Similarly to Section 4.4.1, bootstrapping is employed to estimate confidence intervals for the metrics, providing a robust measure of performance variability, particularly for the small sample size in the paediatric cohort. For context, the results are compared against several weakly-supervised baselines from the previous chapter adapted to the lower-dimensional setting:

1. a 2D-VAE with adversarial training,
2. a 2D-CAM model (Z. Chen et al., 2022), and
3. a transformer-based 2D-WS-MTST model (H. Chen et al., 2023).

In addition, a supervised 2D U-Net pre-trained on the adult dataset is included as an upper performance bound. All baselines are trained and evaluated under identical conditions to ensure a fair comparison.

Important notes

Baseline methods are evaluated only in the pre-trained setting without any fine-tuning. The same results are shown across all subplots to enable direct visual comparison with the 2D-LDM under different configurations.

5.2.2 Pre-trained segmentation model

Evaluation of the pre-trained LDM on paediatric subjects demonstrates that the model generalises effectively across populations, achieving high segmentation performance without any additional fine-tuning. As shown in the first row of Fig. 5.1, high DSC scores are obtained across selected slices and subjects, with performance peaking at 0.6483 for $N = 600, C = 4$. A consistent trend is observed across guidance strengths and sampling schedules, with performance increasing up to $N = 600$ steps, followed by a drop-off beyond this point. Classifier-guidance strength C appears to be influenced by the number of encoding steps, with higher values of C preferred at lower

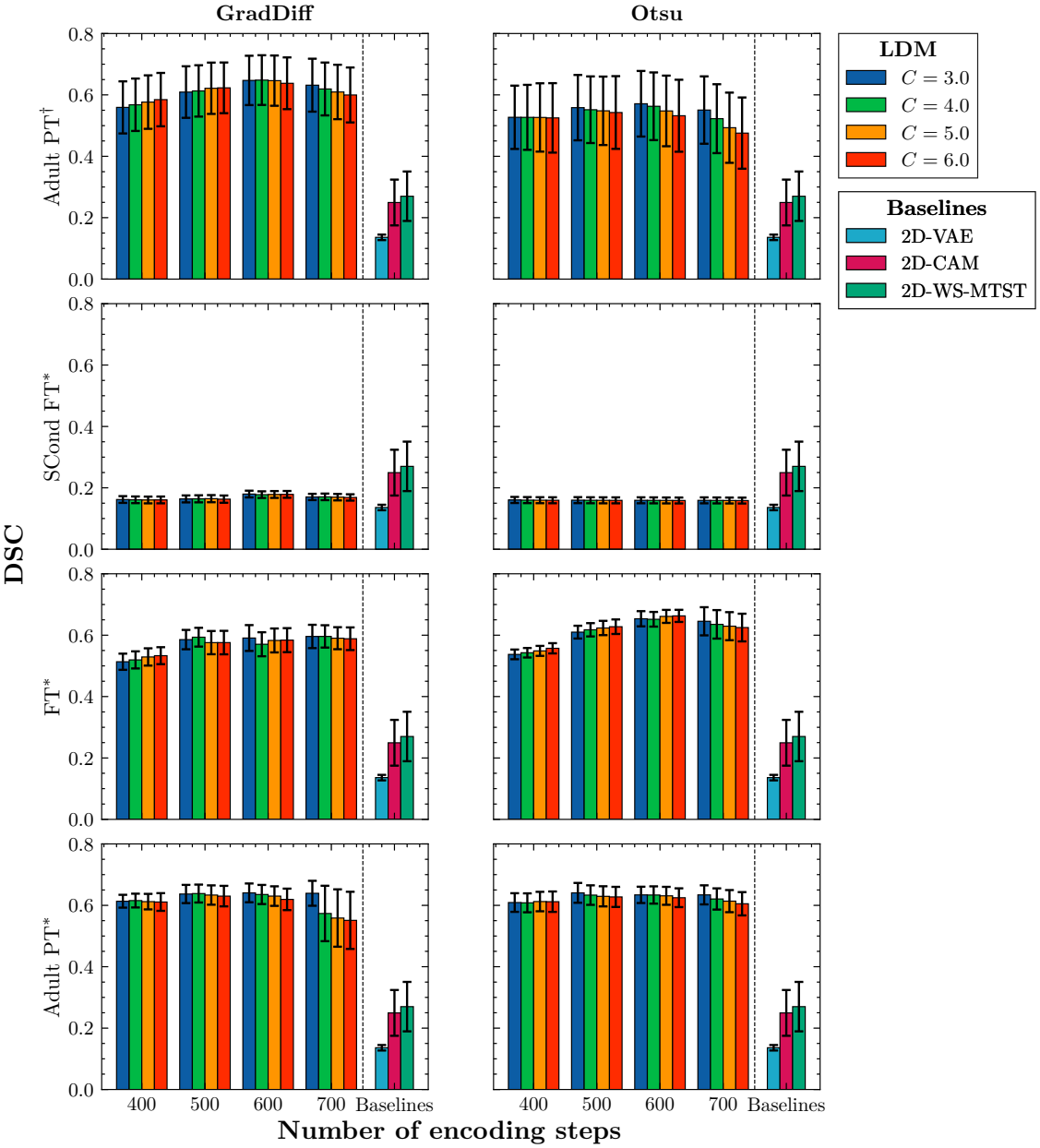


Figure 5.1: Paediatric brain tumour segmentation performance measured by DSC. Each row represents a different training strategy of the 2D-LDM, and each column corresponds to a different thresholding method. The x-axis shows the number of encoding steps N, while the y-axis indicates the DSC score. Colours denote classifier-guidance strength C, and error bars reflect bootstrapped confidence intervals. Baselines are evaluated without fine-tuning and are shown across all plots for visual comparability. PT: pre-trained model; FT: fine-tuned model; SCond: structural encoder conditioning. Superscripts indicate the evaluation set: † refers to the full paediatric test set ($n = 99$), and * refers to the reduced fine-tuning test subset ($n = 20$).

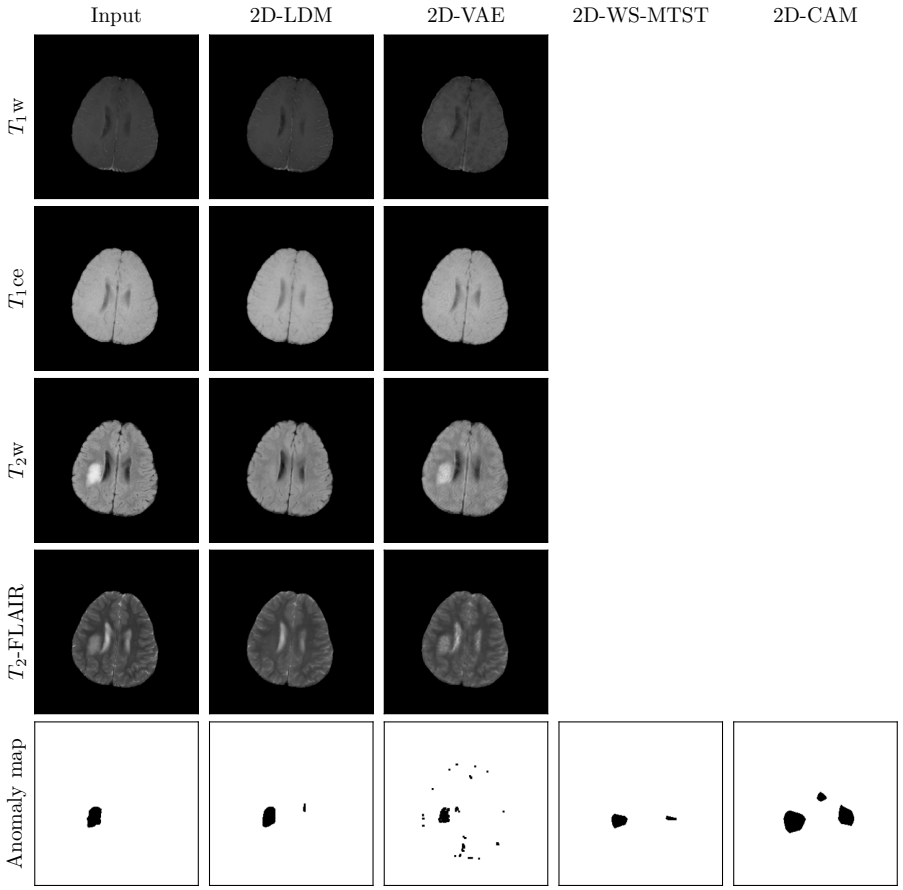


Figure 5.2: Visual results of paediatric brain tumour segmentation. Columns represented the input data followed by all tested configurations. The first four rows represent the MRI sequences and the respective reconstructions or healthy counterfactuals for each model. As 2D-CAM and 2D-WS-MTST are not reconstruction-based or generative models, these sequences are not displayed. The last row indicates the ground truth annotation mask in the first column followed by the model-specific anomaly maps. Results for the 2D-LDM are obtained using the best configuration($N = 600$ and $C = 4$).

N and vice versa. This behaviour reflects the trade-off between guidance strength and sampling steps. Larger C initially improves segmentation by promoting more complete outputs, but also increases the risk of artefacts, as discussed in Chapter 3. At higher N, these risks outweigh the benefits, explaining the observed performance drop-off. Interestingly, the variance appears to be unaffected by any hyperparameter, suggesting that the model's uncertainty is more influenced by inherent data variability than by sampling configurations.

GradDiff consistently outperforms Otsu thresholding across all configurations, achieving higher DSC scores with reduced variance. Designed to be sensitive to subtle intensity differences (see Section 4.4.1), GradDiff appears to be more effective in mitigating the impact of distributional shifts between adult and paediatric data, leading to 5 % to 10 % improvements in DSC.

Similar trends are reflected in specificity, reported in Fig. C.2, which remains stable up to $N = 600$ before degrading with further diffusion steps. The same preference for lower classifier guidance strength is evident, again with $C \in [3, 4]$ producing the most favourable outcomes. Compared to Otsu thresholding, GradDiff achieves higher specificity and more compact result distributions, suggesting improved robustness and reduced susceptibility to false positives.

Compared to the baseline models, the pre-trained 2D-LDM achieves substantially higher performance. The reported DSC scores of the baselines are consistent with the findings in Chapters 3 and 4, with the VAE performing lowest at 0.1360, followed by 2D-CAM at 0.2495 and its advanced variant, 2D-WS-MTST, at 0.2670. The notably reduced performance is visualised in Fig. 5.2, where the pre-trained 2D-LDM with $N = 600$ and $C = 4$ almost perfectly detects the brain tumour in the paediatric subject, while the baselines either produce an abundance of outliers (2D-VAE) or detect additional larger lesions within the same subject (see the 2D-CAM and 2D-WS-MTST results). In comparison, the fully supervised U-Net shown in Fig. C.1 attains a DSC of 0.9480, establishing a strong upper bound for this cohort. Although the 2D-LDM does not reach this level, it clearly outperforms every other tested model using only image-scale annotations.

5.2.3 Fine-tuning adult model

This section investigates fine-tuning strategies applied to the pre-trained model of the previous section with the aim of extending its utility to paediatric data. The first strategy, outlined in Section 5.2.3, leverages the structural conditioning mechanism described in Section 2.3.3 to adapt the pre-trained model to the target task. This choice is motivated by the preceding chapter, where adaptation through conditioning was shown to be effective for both controlled synthesis and resolution enhancement (see Chapter 4). In parallel, classical fine-tuning was explored by adapting the parameters of the pre-trained model using target task data (see Section 5.2.3). Full re-training was not pursued, as the objective of this chapter is to assess the generalisability of weakly-supervised learning rather than to develop task-specific models. Moreover, given the limited sample size and the high capacity of the underlying DDPM, training from scratch would likely result in overfitting, thereby undermining the intended evaluation (Ayana et al., 2024; Ganatra, 2025).

Fine-tuning using encoder conditioning

In the structurally conditioned variant, fine-tuning is performed by introducing additional learnable parameters through a structural encoder. This encoder follows the design of the U-Net encoder but constitutes the only component updated during training, while the backbone weights remain frozen. In this way, spatial context is injected into the backbone U-Net without altering its pre-trained architecture. Given its prior success in conditional synthesis and SR, this strategy was considered a viable alternative for extending model utility and further assessing the generalisability of the weakly-supervised formulation.

Fine-tuning with the structural encoder resulted in a substantial degradation of performance relative to the pre-trained baseline, as shown in Fig. 5.1, second row. Across most configurations, the DSC score dropped by more than 40%, with the best performance reaching only 0.1782. Despite the overall decline, the similar performance trends to the non-pre-trained model were observed: performance peaked

at $N = 600$, and conditioning via GradDiff yielded marginal improvements over alternatives. These results indicate that the inclusion of the structural encoder does not enhance segmentation performance in this setting.

One possible explanation lies in the nature of the conditioning mechanism. In the previous chapter, the structural encoder successfully adapted a pre-trained LDM to novel tasks such as synthesis and SR, where the conditioning inputs introduced distinct and complementary information. In contrast, the current task does not deviate substantially from the original training objective. As a result, the additional conditioning signal may introduce redundancy or even interfere with learned representations, leading to degraded performance.

Another possible explanation lies in overfitting to the limited amount of data available. Chapter 4 provided roughly 10-15 times more training samples. Under those conditions the structural encoder showed clear benefits of re-utilising pre-trained priors. Both explanations remain difficult to disentangle, but together they suggest that structural conditioning is most effective when adequate data are available and when the target task introduces additional context not already captured during pre-training.

Classical fine-tuning

Classical fine-tuning adapts the pre-trained adult model to the paediatric domain by updating its layers on the target data. In this setup, the pre-training serves primarily as a prior to accelerate convergence. The core objective remains unchanged, and the model continues to operate within the same weakly-supervised anomaly detection framework. This approach capitalises on the shared structure of the tasks while avoiding the need to initialise training from random weights.

As shown in Fig. 5.1, third row, fine-tuning affects the two thresholding mechanisms differently. For GradDiff, performance decreased across all tested configurations, with reductions of 1 % to 7 % relative to the pre-trained model. The largest drop occurred at $N = 600$. In contrast, Otsu thresholding benefited from fine-tuning, achieving gains of 1 % to 15 %, particularly at high N and C. The best result was

obtained with $N = 600$ and $C = 6$, yielding a DSC of 0.6629 , slightly surpassing the best pre-trained configuration. These results partially restores the previously observed ratio between the two thresholding strategies Section 4.4, where GradDiff shows inferior performance to Otsu thresholding.

Both thresholding methods also exhibited reduced variance, which can be attributed due to the smaller test set size used during fine-tuning (see Section 5.2.1). The latter is confirmed by evaluating the pre-trained model of Section 5.2.2 on the reduced test set, resulting in comparably low variances (see Fig. 5.1, bottom row).

These results suggest that GradDiff performs competitively without fine-tuning, achieving scores comparable to fine-tuned Otsu thresholding. This highlights the strength of gradient-based thresholding as a mechanism for guiding anomaly localisation, particularly under distributional shifts where explicit re-training may be impractical or suboptimal.

In addition, these results further validate the hypothesis outlined in Section 5.2.3, namely that the structural encoder is only effective when the target task diverges meaningfully from the pre-training objective and sufficient data are available to support adaptation. Given the limited size of the paediatric dataset, classical fine-tuning proves more robust, as it preserves or improves performance under the same data constraints. This can be attributed to the fact that the pre-trained model already captures relevant features for the paediatric domain, requiring only minor adjustments rather than the introduction of new parameters. As a result, less data are needed to adapt the existing representations effectively, compared to the structural encoder, which effectively re-initialises a large portion of the model.

5.2.4 Summary of findings

The adult-pretrained 2D-LDM demonstrates clear advantages over conventional weakly-supervised baselines, achieving substantially higher segmentation performance and accommodating the adult-to-paediatric distribution shift more effectively. In contrast, baseline models such as the 2D-VAE, 2D-CAM, and 2D-WS-MTST show limited transferability, reflecting their reduced capacity to capture complex anatomical

and pathological variation. When compared to a fully supervised U-Net, the LDM remains competitive despite relying only on weak supervision. This positions the model at the same level of robustness as supervised approaches under mild distribution shifts. This capacity is attributable to its encoding mechanism, which preserves global anatomical structure while enabling conditional edits in relevant regions.

Fine-tuning strategies provide only minor improvements, with classical fine-tuning yielding modest gains for specific configurations, while structural encoder-based fine-tuning proves ineffective. These findings indicate that the pretrained model already captures transferable representations, with adaptation offering limited additional benefit. Overall, the results underscore the generalisability of the weakly-supervised DDPM-based framework to paediatric cohorts and highlight its potential as an annotation-efficient alternative to fully supervised methods in data-scarce clinical scenarios.

5.3 Evaluation on private data collection

The findings of the previous section demonstrate the generalisability of the proposed weakly-supervised LDM to paediatric populations using a publicly available benchmark dataset. However, real-world clinical data pose additional challenges, including variability in acquisition protocols, scanner hardware, and patient demographics. To assess the robustness and applicability of the framework under these more realistic conditions, a complementary clinical dataset was assembled through collaborations with The Children's Hospital at Westmead (CHW), Australia, and The Children's Hospital of Philadelphia (CHOP), United States of America. The dataset forms a diverse resource for examining cross-institutional performance and the practical utility of the proposed approach in paediatric neuro-oncology.

5.3.1 Description of dataset

Both datasets were acquired under appropriate ethical approvals and in compliance with data protection and patient confidentiality standards. The CHW dataset was

curated locally from existing records and therefore entails an in-depth description of the data curation process. In contrast, the CHOP dataset was obtained through The Children's Brain Tumor Network (CBTN), which provides de-identified research data to external investigators upon request. The following sections describe both datasets in detail, including their characteristics and the procedures employed to ensure ethical compliance and data integrity.

CHW dataset

The first part of the complementary dataset was obtained from the CHW and is designed as a healthy control group for model fine-tuning and adaptation. It consists of 151 MRI acquisitions from anatomically healthy children and adolescents aged between 3 and 18 years. The collection encompasses retrospectively gathered MRI scans, acquired between January 2008 and July 2023. The children were administered for neurological symptoms without any anatomical changes observed during MRI examination. This collection provides a representative healthy control group spanning a wide paediatric age range. The CHW dataset also includes a separate pre-existing cohort of diffuse midline glioma (DMG) patients obtained from a pre-existing study (2020/STE03292). The project presented in this chapter, including both dataset components, is formally registered as "Analysis of paediatric brain tumours using Machine Learning" (2023/ETH01816) on the Research Ethics Governance Information System and holds site-specific approval from CHW (2023/STE03408). Ethical approval was also obtained from the human research ethics committee (HREC) of The University of Technology Sydney (UTS) (ETH24-9548).

Dataset curation process Due to the ethical sensitivities and privacy risks associated with paediatric imaging data, all procedures were conducted in strict compliance with institutional and state regulations, as outlined in Section 2.1.1. To facilitate access under these conditions, I obtained formal affiliation with CHW as a contingent worker. Eligible participants were identified from the Cancer Centre for Children departmental database, and requests for de-identified imaging were

submitted to the Medical Imaging Department. The selection of the healthy control group was independently performed by three clinical specialists, who reviewed both the imaging data and corresponding reports to confirm the absence of pathology. All curation procedures were conducted in accordance with National Health and Medical Research Council recommendations and New South Wales privacy legislation.

Serial MRI acquisitions in Digital Imaging and Communications in Medicine (DICOM) format were collated by the research team at CHW. The process was supervised by the principal investigator (PI), Dr. Robert Goetti, and coordinating PI, Prof. Daniel Catchpoole, who ensured that de-identification was complete prior to transfer. All identifying DICOM fields (e.g. patient name, date of birth, and medical record number) were removed using scanner software at the time of export. Each case was assigned a study number, with the re-identification key retained securely within the hospital firewalls by the PI. The de-identified imaging data were encrypted, and transferred to UTS by the coordinating PI, where they are stored on a password-protected server, with exclusive access to the research team listed on the ethics application. The encrypted hard drive was then returned to CHW. No identifiable data were transmitted outside the hospital.

The data curation process extended over more than a year, beginning in mid-2022 with initial applications and feasibility discussions with medical professionals, and concluding with final approval in October 2023. This early start was essential to secure the multiple ethics approvals required for paediatric data access. At the same time, the pipeline was still under development, and its full capacities and limitations were not yet known at the point of application.

Important note

Although this dataset was curated with considerable effort over more than a year, it could not be utilised in the final experiments. The absence of the full set of MRI sequences required for the 2D-LDM trained on the BraTS dataset (Section 2.1.1) made both the diseased and healthy cohorts incompatible with the finalised pipeline. Moreover, as shown in Section 5.2, fine-tuning strategies did not yield improvements, further reducing the value of the healthy control cohort in the present setting. The dataset is therefore included here primarily to illustrate the practical challenges and extensive efforts involved in obtaining paediatric imaging data for research.

CHOP dataset

While the locally curated dataset could not be employed in the final experiments, complementary data were available from CHOP through the CBTN, providing a suitable basis for evaluation. This research dataset is accessible to investigators upon request and provides a diverse collection of paediatric brain tumour cases. It was included both for its heterogeneity and because it originates from a different institution, enabling the assessment of generalisability across cancer types as well as across acquisition sites. The dataset encompasses a wide spectrum of paediatric brain tumour variants, including LGGs, HGGs, ependymoma, and DMG cases. This diversity makes it possible to evaluate the proposed anomaly detection framework on tumour types commonly encountered in paediatric neuro-oncology (see Section 1.1), and to identify both its strengths and its limitations in handling clinically relevant heterogeneity.

In addition, the dataset comprises multiple MRI sequences for each case, including T_1-weighted (T_1w), T_1ce, T_2-weighted (T_2w), and T_2-fluid attenuated inversion recovery (T_2-FLAIR) modalities. These sequences are required for the model designed in this thesis, which relies on the same multi-modal input structure as the BraTS dataset. As a result, each tumour subgroup was analysed based on the availability of these sequences and processed using the same standardisation pipeline applied

to the BraTS dataset. This pipeline is provided by the Cancer Imaging Phenomics Toolkit[1] (Davatzikos et al., 2018; Pati et al., 2020), which performs co-registration to a common template, skull-stripping, bias field correction and resampling to an isotropic resolution of $1\,\mathrm{mm}^3$. In addition, the pipeline features an automated segmentation module based on a pre-trained 3D U-Net, which obtains a preliminary tumour mask for quantitative and qualitative assessment.

The resulting dataset comprises the following statistics per tumour type:

- **Ependymoma**: 189 individual acquisitions from 189 individuals
- **LGG**: 505 individual acquisitions from 505 individuals
- **HGG**: 263 individual acquisitions from 253 individuals
- **DMG**: 65 individual acquisitions from 65 individuals

5.3.2 Experimental results

This section presents preliminary results obtained by applying the pre-trained model from Section 5.2.2 to the curated clinical dataset described in Section 5.3.1.

Preprocessing and evaluation

As outlined in Section 5.3.1, only the CHOP dataset was used for evaluation due to the incompatibility of the CHW dataset. Since the CHOP dataset was processed with the same pipeline as the BraTS dataset, it is directly compatible with the pretrained model from Section 5.2.2. Additionally, it follows the identical preprocessing, inference, and evaluation procedures described in Section 5.2.1. This ensures consistency with previous experiments and allows direct comparison of results. Preprocessing following the BraTS processing pipelines includes padding to a base-2 resolution of 256×256 pixels and normalisation to the range $[-1, 1]$.

Inference uses the same encoding and classifier-free guidance strategy, with segmentation masks derived from the difference between the original and counterfactual images. Both Otsu and GradDiff thresholding methods are applied to isolate lesion regions. Because the available annotations were generated automatically by the

[1] https://cbica.github.io/CaPTk/preprocessing_brats.html

BraTS pipeline (see Section 5.3.1) rather than confirmed by clinical experts, the quantitative analysis should be interpreted with caution.

For subgroup-specific assessment, the dataset is partitioned into four tumour categories (ependymoma, LGG, HGG, and DMG). From each subgroup, 50 individuals with two lesion-containing slices are selected, yielding 100 slices per tumour type. This sampling strategy introduces variability while accounting for the fact that the automated segmentation used in the BraTS preprocessing pipeline may not detect lesions in all subjects. The buffer ensures balanced selection across subgroups, including DMG, which comprises only 65 cases in total. As in the previous chapter, performance is estimated through bootstrapping, drawing resampled sets from the available data to provide more robust estimates of variability. Evaluation combines preliminary quantitative metrics with visual inspection of segmentation quality, with particular attention to edge cases and clinically relevant scenarios. Reported metrics include DSC and specificity, though their reliability is limited by the absence of expert-confirmed annotations.

Experimental results on CHOP dataset

As outlined before, the results are split into quantitative and qualitative analyses. Quantitative metrics are presented first to provide an overview of performance across the entire dataset, followed by a detailed qualitative assessment of representative cases to illustrate the model's strengths and limitations in practical scenarios.

Quantitative analysis Across tumour subgroups, the 2D-LDM exhibits trends consistent with those observed in the paediatric BraTS dataset. For HGG, performance remains relatively stable, reaching a best DSC of 0.4172 with $N = 500$, $C = 3$, and Otsu thresholding. The corresponding configuration with GradDiff yields 0.3896, reflecting the typical 3 % to 4 % reduction observed throughout Fig. 5.3. A similar pattern is evident for ependymoma, where performance is lower overall but stable, achieving 0.2733 with Otsu and 0.2003 with GradDiff, corresponding to a decrease of approximately 7 %.

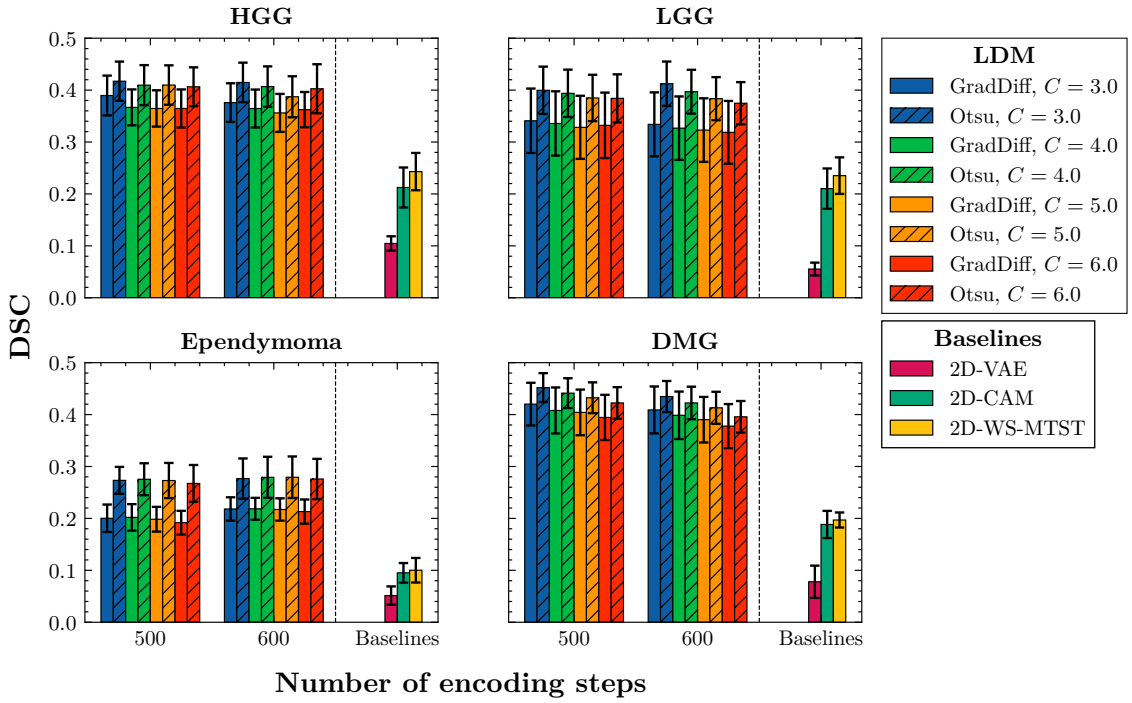


Figure 5.3: Quantitative results for the CHOP dataset. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis reports the DSC. Each subplot corresponds to a tumour subtype. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

In contrast, LGG and DMG show sensitivity to the gradient scale C, consistent with findings in Section 5.2. For LGG, the DSC declines by nearly 2 % between $C = 3$ and $C = 6$ across both thresholding methods. This effect is amplified in DMG, where performance also depends strongly on N, resulting in a further reduction of almost 3 %. Notably, DMG attains the highest overall score among subgroups, despite not being represented during training, with a best DSC of 0.4520 ($N = 500$, $C = 3$, Otsu). The corresponding GradDiff score is 0.4202, approximately 3 % lower.

When compared to other weakly-supervised baselines, the 2D-LDM achieves substantially higher average performance on this dataset: 25 % for DMG, 18 % for LGG, and 17 % for both HGG and ependymoma relative to the best-performing 2D-WS-MTST. Only the supervised U-Net surpasses these results (see Fig. C.3), although the margin is considerably smaller than on the BraTS dataset (see Fig. C.1).

Specificities are high across all methods, with the 2D-LDM only about 1 % lower than the baselines, as shown in Fig. C.4. This minor difference reflects the tendency of the baselines to default to background predictions, which inflates specificity but comes at the cost of markedly lower DSC.

In summary, the 2D-LDM sustains its performance across previously unseen tumour types and acquisition conditions, reinforcing its generalisability. Nevertheless, segmentation accuracy remains highly dependent on hyperparameters, and the evaluation itself is constrained by reliance on the pre-trained U-Net from the BraTS pipeline, which shows difficulties in edge cases (e.g. first row of Fig. 5.4). This complicates the assessment of thresholding strategies and may partially contribute to the reduced performance of GradDiff in this setting.

Qualitative analysis Representative qualitative results are shown in Fig. 5.4, illustrating the variability in prediction quality across subjects. In several cases (rows 2, 4, 5, 6, 7, and 8), the model generates plausible healthy counterfactuals, while in others the reconstructions remain inaccurate. This variability cannot be explained by lesion size alone. For instance, the large lesion in row 1 proved challenging to remove, but the equally large lesion in row 5 was mostly corrected. Moreover, the difficulty in row 1 is not unique to the proposed approach; the pre-trained 3D U-Net from the BraTS preprocessing pipeline also fails to delineate the anomaly accurately (Fig. 5.4c).

Another notable observation is the 2D-LDM's ability to operate on scans with limited fields of view, as illustrated in row 3 of Fig. 5.4. Here, the brain is not fully captured, yet the model refrains from introducing spurious anatomy into the missing regions and instead focuses on reconstructing the structures present. This behaviour supports the effectiveness of the encoding mechanism, which preserves intact anatomy and confines conditional edits to the relevant regions.

The approach also generalises to out-of-distribution tumour types. Accurate counterfactuals were obtained for ependymoma (rows 5 and 6) and DMG (rows 7 and 8), closely resembling the predictions of the supervised U-Net ground truth.

Figure 5.4: Visual results for the CHOP dataset. (a) shows the input data; (b) presents the healthy counterfactual generated by the best-performing configuration of the pre-trained 2D-LDM ($N = 600$, $C = 3$); (c) displays the ground-truth annotation; (d) and (e) show the anomaly maps obtained using Otsu and GradDiff thresholding, respectively. Rows correspond to different subjects, with every two subjects representing the same tumour type. From top to bottom:HGG, LGG, ependymoma and DMG. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

This further underscores the generalisability of the framework, which successfully detects non-normal tissue despite being trained exclusively on healthy, LGG, and HGG cases. Any preferences to a specific tumour type for the present samples are not evident beyond the aforementioned size effect, which primarily affects HGG cases due to their aggressive nature and growth.

Finally, artefacts introduced during generation can offset the anomaly map calculation as observed throughout this thesis. These artefacts could not be mitigated by any of the methods evaluated in this study and directly reduce detection reliability. This challenge was partially addressed in Chapter 4 through the use of SR, which improved small lesion detection but was not pursued here due to the considerable computational cost associated with high-resolution training and inference (see Section 4.4.4). As illustrated in Fig. 5.4e, GradDiff thresholding can produce more compact segmentations than Otsu, yet these do not overcome the fundamental limitations imposed by artefacts in the generated images.

5.4 Limitations and conclusion

This chapter has demonstrated the generalisability of the proposed weakly-supervised framework to paediatric populations by leveraging pre-trained 2D-LDMs and their ability to transfer across cohorts. The pre-trained 2D-LDM achieved performance comparable to the fine-tuned variant, supported by the novel GradDiff thresholding method, which proved more robust and consistent than conventional thresholding in this setting. The framework further surpassed conventional weakly-supervised baselines, achieving substantially higher segmentation performance and accommodating the adult-to-paediatric distribution shift more effectively. This capacity is attributable to the encoder design, which preserves global anatomical structure while enabling conditional edits in relevant regions, thereby supporting robust generalisability under distributional change.

Despite outperforming other weakly-supervised methods, the results on the paediatric BraTS dataset also show that the supervised baseline generalises strongly.

This may suggest that the adult-to-paediatric shift is relatively modest; at least in this controlled benchmark: sufficient to challenge conventional weakly-supervised baselines but not severe enough to establish whether weakly-supervised anomaly detection with LDMs is inherently better suited to handling distribution shifts than supervised deep learning (DL) models. The preliminary analysis of the CHOP dataset provides a different perspective. Here, the 2D-LDM generalises beyond the tumour types represented in BraTS, achieving comparable performance in some cases to the official 3D U-Net used in the BraTS preprocessing pipeline. These results indicate that the distributional differences between adult and paediatric data are more substantial than the BraTS-only evaluation suggests, and that the weakly-supervised 2D-LDM may be more robust to such shifts. At the same time, the analysis exposes a critical limitation: reliance on complete multi-sequence inputs and highly curated preprocessing pipelines reduces flexibility when faced with real-world clinical data, which often exhibit missing modalities, variable acquisition protocols, and inconsistent coverage. While the proposed encoding mechanism preserves available anatomy, further investigation is needed to develop approaches that operate reliably under these less controlled conditions. An additional limitation is that the quantitative evaluation relies on the automatic annotations provided by the BraTS preprocessing pipeline, which can be inaccurate in challenging cases. A more definitive assessment will therefore require larger-scale studies spanning multiple diseases and populations, supported by expert-validated annotations, where variability is more pronounced and quantitative evaluation more reliable.

A further limitation arises from the rigidity of the proposed approach. Since the models are trained on the full set of complementary MRI sequences provided by the BraTS dataset, they also depend on these sequences during inference. In practice, many clinical datasets, particularly retrospective collections, lack complete multi-sequence coverage, as demonstrated by the CHW cohort analysed in this chapter. This dependency reduces the generalisability of the method and restricts its applicability, because individuals with incomplete data are excluded from evaluation. Another constraint is that the results in this chapter were obtained using the original

hyperparameter configuration of GradDiff defined in Section 4.4.1. As shown in the ablation study of SR (see Section 4.4.3), adapting these parameters can yield improved outcomes in certain settings. The findings here highlight mixed behaviour across datasets, suggesting that additional tuning of GradDiff may be necessary to realise its full potential. Nevertheless, expanding the parameter space increases the overall complexity of the method, and the consistency of such gains across different datasets remains uncertain.

Lastly, a limitation arises from the fine-tuning strategy employed for adapting the pre-trained LDM to paediatric data. This work adopted the encoder-based fine-tuning approach proposed by J. Wang et al. (2024), which proved highly effective for synthetic image generation and SR tasks in Chapter 4. However, its application in this context yielded reduced performance compared to classical fine-tuning, which achieved parity with the pre-trained model. One possible explanation is that encoder-based fine-tuning is most effective when the target task diverges substantially from the original training objective. In contrast, anomaly detection on paediatric data remains closely aligned with the adult training domain, limiting the benefit of decoupling the encoder and potentially introducing unnecessary adaptation noise. While it would be desirable to integrate disease- or population-specific encoders for targeted adaptation, this strategy did not prove effective here. Future work may therefore require exploration of alternative fine-tuning methods, unless the generalisability observed in this study extends beyond the adult-to-paediatric shift.

In conclusion, this chapter has successfully demonstrated the generalisability of the proposed weakly-supervised anomaly detection framework to paediatric brain tumours. As a result, the proposed approach address the third research gap outlined in Section 2.4.3. In summary, I:

- Assessed the generalisability of the weakly-supervised framework to paediatric populations, leveraging pre-trained models and their capacity to generalise to paediatric brain tumours.
- Demonstrated that the pre-trained 2D-LDM achieves comparable performance to a fine-tuned model, owing to the novel thresholding method GradDiff.

- Highlighted the limitations of the encoder-based fine-tuning strategy, which proved less effective than classical fine-tuning in this context.
- Curated a real-world dataset through collaborations with multiple institutions, evaluated the proposed approach under realistic clinical conditions, and highlighted key challenges of unprocessed clinical data.

These efforts culminate in the following contribution:

Contribution 4

Demonstrated the generalisability of DDPM-based weakly-supervised anomaly detection for brain tumour segmentation to data-scarce paediatric cases, validated on a curated multi-institutional dataset under clinically realistic conditions.

Chapter 6

Conclusion

Paediatric brain tumours remain a major global health concern, particularly due to the complexity and severity of their clinical implications. Accurate identification and delineation of tumoural lesions are central to diagnosis, treatment planning, and ongoing monitoring. Ensuring the highest possible accuracy and reliability in this process is therefore of critical importance. Computer-aided diagnosis (CAD) systems have the potential to provide valuable guidance and support to clinicians by enhancing the consistency and sensitivity of medical image analysis, particularly in cases where human interpretation may be limited by ambiguous presentations.

The overarching aim of this thesis was to reduce reliance on costly manual annotations for brain tumour segmentation while narrowing the performance gap between supervised and weakly-supervised deep learning (DL) models. In Chapter 3, I addressed Gap 1 by proposing a patch-based sampling strategy that extracts pseudo-healthy subvolumes from diseased individuals, enabling effective training of a 3D-latent diffusion model (LDM) under limited supervision. A more robust encoding mechanism was also introduced to preserve healthy anatomy and support anatomically plausible reconstructions, leading to improved segmentation performance. In Chapter 4, I addressed Gap 2 by constructing a synthetic dataset of small brain tumours to enable controlled evaluation, and by examining how super-resolution (SR) enhances lesion visibility and detection sensitivity using a unified framework. Finally, in Chapter 5, I addressed Gap 3 by evaluating generalisability to paediatric data, where scarcity and anatomical variability present further challenges. Leveraging a

curated paediatric dataset, I demonstrated that pretrained LDMs can adapt with limited modifications and remain robust to mild distribution shifts. Collectively, these contributions advance weakly-supervised brain tumour segmentation with denoising diffusion probabilistic models (DDPMs):

Contribution 1 Developed a patch-based LDM for efficient 3D weakly-supervised brain tumour segmentation via anomaly detection, preserving volumetric context and integrating a robust encoding strategy to improve anatomical fidelity.

Contribution 2 Developed and validated a DDPM-based model for the controlled synthesis of size-specific brain tumour lesions in multi-sequence magnetic resonance imaging (MRI).

Contribution 3 Conducted the first systematic investigation of DDPM-based SR in weakly-supervised anomaly detection, evaluating its impact on the detectability of small brain tumours.

Contribution 4 Demonstrated the generalisability of DDPM-based weakly-supervised anomaly detection for brain tumour segmentation to data-scarce paediatric cases, validated on a curated multi-institutional dataset under clinically realistic conditions.

The remainder of this chapter is now structured as follows: Section 6.1 offers a detailed reflection on the contributions of this thesis, linking them to the research questions and gaps identified in Section 2.4, and summarising their significance and limitations. Section 6.2 discusses potential directions for future research that emerge directly from this work. Finally, Section 6.3 closes the chapter with a broader perspective on the clinical relevance and long-term impact of this research in advancing medical imaging and diagnostic support.

6.1 Reflection on contributions

Contribution 1

Developed a patch-based LDM for efficient 3D weakly-supervised brain tumour segmentation via anomaly detection, preserving volumetric context and integrating a robust encoding strategy to improve anatomical fidelity.

Contribution 1 addressed Research Question 1 (*How can weakly-supervised 3D anomaly detection be made computationally efficient while mitigating reliance on publicly available datasets of healthy individuals?*) by focusing on Gap 1 (*The transition from 2D to 3D medical image analysis for weakly-supervised brain tumour segmentation remains underexplored due to increased computational demands and the lack of public datasets with healthy individuals.*).

Chapter 3 introduced and evaluated a novel 3D-LDM designed to address the computational constraints associated with volumetric weakly-supervised segmentation. By transitioning from full-resolution input to a condensed latent space representation, I substantially reduced the memory requirements and improved the efficiency of 3D operations. To further support this design, I developed a tailored patch-based sampling mechanism that further reduced the computational cost and enabled the extraction of pseudo-healthy regions from pathological scans. This approach leveraged the plasticity of the human brain, assuming that regions distant from lesions retain structurally healthy features. The hypothesis was validated through improved segmentation performance when compared to uniform sampling, demonstrating that even diseased datasets can serve as a viable resource for anomaly detection.

In extensive benchmarking, the proposed 3D-LDM outperformed all state-of-the-art weakly-supervised methods, achieving higher accuracy while offering faster and more stable training and inference. These results highlight the model's ability to focus on clinically relevant features under weak supervision. Furthermore, the choice of encoding mechanism was shown to be a critical design factor. The integration of exact diffusion inversion via coupled transformations (EDICT) facilitated stronger

conditional edits with minimal impact on healthy anatomy, and improved inference speed by allowing shorter diffusion sequences without degrading performance.

Despite these advances, the framework is not without limitations. Artefact generation during classifier-guided inference remains a prominent challenge, particularly in background regions or anatomically sparse areas. These artefacts affect the reliability of intensity-based thresholding by shifting value distributions and increasing false negatives. The model's sensitivity to classifier strength and number of encoding steps further complicates the tuning process. While higher guidance strengths can improve segmentation, they may also introduce instability and exacerbate artefacts. Similarly, while shorter encoding sequences accelerate inference, they require careful balancing to preserve performance and mitigate noise. Inconsistent outcomes across visually similar cases suggest that robustness remains an area for improvement, particularly where inaccurate lesion removal compromises segmentation reliability.

Nonetheless, this contribution represents a substantial step forward for weakly-supervised brain tumour segmentation using DDPMs. It directly addresses the challenge outlined in Gap 1 and answers the associated Research Question 1, providing an efficient and effective 3D framework. Most importantly, it established the viability of LDMs as modular framework to detect anomalous regions in medical images. These findings laid the foundation for subsequent investigations into small lesion detectability, where thresholding may be particularly fragile, and cross-domain generalisability to paediatric cohorts, where data scarcity and anatomical variability present further challenges.

Contribution 2

Developed and validated a DDPM-based model for the controlled synthesis of size-specific brain tumour lesions in multi-sequence MRI.

Contribution 3

Conducted the first systematic investigation of DDPM-based SR in weakly-supervised anomaly detection, evaluating its impact on the detectability of small brain tumours.

Contribution 2 and Contribution 3 addressed Research Question 2 (*What is the effect of DDPM-based SR on the sensitivity and segmentation performance of weakly-supervised anomaly detection, particularly for small brain tumours?*) by focusing on Gap 2 (*The role of SR in enhancing weakly-supervised DDPM-based anomaly detection remains underexplored, particularly with respect to its effect on sensitivity to small or subtle brain tumours.*).

Chapter 4 investigated the use of conditioning mechanisms in LDMs to support two interconnected yet distinct tasks: (1) the generation, and (2) the detection of small brain tumours. To facilitate the detection of small lesions under controlled conditions, I first developed a method to synthetically generate datasets that exhibit specific lesion characteristics. This required a comprehensive analysis of conditioning strategies, given that the replication of established approaches was met with limited success. The findings revealed that the LDM framework is particularly responsive to binary mask-based conditioning, enabling the generation of anatomically coherent datasets tailored to the detection task. These generated samples served as a foundation to explore the limits of lesion detectability.

In an effort to investigate the capacity for early detection of small lesions, I focused on the worst-case scenario of detecting faint distributional differences under weak supervision. To improve the framework's sensitivity, I extended the thresholding mechanism with gradient-based information to enhance both sensitivity and specificity in challenging cases. Building on this foundation, I employed SR to increase the effective resolution of the input, a strategy that proved particularly valuable for lesions with minimal spatial extent, where intensity differences are subtle and easily obscured by noise. The results confirmed that higher resolution improved detectability and reduced variance across subjects, while consistently outperforming state-of-the-art weakly-supervised approaches that do not employ DDPMs for anomaly detection. Additionally, the adapted degradation pipeline incorporating more clinically relevant artefacts showed no substantial differences to the established method, but it demonstrates the flexibility of the approach and its potential for further adaptation to other datasets.

However, the generative component presented several limitations. Generating realistic synthetic data proved more complex than anticipated due to the influence of the chosen conditioning mechanism and conditioning signal. Simple conditioning mechanisms such as latent-space concatenation, as well as advanced strategies re-using pre-trained models, produced visually and quantitatively compelling results. Nevertheless, subtle distribution shifts introduced by these methods can impair downstream models, underscoring the need for careful evaluation beyond sample quality and including task-specific performance. In addition, the binary lesion masks used for conditioning proved simple yet effective for normal-sized lesions. Nevertheless, the limited information content of this small signal relative to the larger background suggests the need for alternative conditioning strategies that remain stable in edge cases involving very small lesions. While the current approach is effective for common lesion appearances, further adaptation will be necessary to ensure robust performance in such challenging scenarios.

On the detection side, inference time emerged as a major bottleneck. The use of SR increased computational time, as patch-based inference with overlap was required to overcome memory constraints. Additionally, attempts to unify conditional SR and healthy counterfactual segmentation into a single process were unsuccessful, necessitating the sequential application of both models and thereby substantially increasing inference time. While the healthy counterfactual generation was shown to be more resilient to distribution shifts, the SR model exhibited performance degradation when applied to the synthetic dataset, pointing to domain mismatch despite synthetic alignment efforts.

Lastly, a limitation arises from the novel gradient-based thresholding mechanism GradDiff. It was designed to better capture subtle differences and showed improved sensitivity over traditional thresholding in initial grid-search experiments on regular lesions. However, this advantage did not fully carry over to the final evaluation set, where the method underperformed relative to conventional approaches. To further examine its potential, an additional ablation study was conducted on small lesions with adapted parameters. This analysis demonstrated that, when carefully

tuned, GradDiff can provide measurable benefits in detecting small lesions, with performance improvements over intensity-based thresholding. It also showed that memory requirements can be reduced by disabling latent gradient computation and relying on classifier-free pseudo-gradients, with only minimal impact on accuracy.

Despite these limitations, this chapter makes meaningful contributions to the field of weakly-supervised brain tumour segmentation with DDPMs. It demonstrates that LDMs can be conditioned to generate datasets with precise lesion characteristics, and that these datasets can be used to evaluate and enhance detection performance under controlled scenarios. The promising results of gradient-based thresholding, together with domain-specific degradation strategies and resolution-aware detection pipelines, mark an important step toward increasing the clinical utility and robustness of weakly-supervised anomaly detection. Moreover, this work highlights the potential of pre-training and encoder-based conditioning to incorporate learned priors into downstream tasks. These components emerged as promising strategies for improving generalisability across data distributions, setting the foundation for the final contribution focused on paediatric brain tumours.

Contribution 4

Demonstrated the generalisability of DDPM-based weakly-supervised anomaly detection for brain tumour segmentation to data-scarce paediatric cases, validated on a curated multi-institutional dataset under clinically realistic conditions.

Contribution 4 addressed Research Question 3 (*To what extent can DDPMs trained on adult brain tumour data generalise to the paediatric domain, and how robust are the learned representations to shifts in population and disease distribution?*) by focusing on Gap 3 (*The generalisability of weakly-supervised DDPMs to paediatric brain tumours remains untested, despite their hypothesised robustness to distributional shifts and suitability for data-scarce clinical settings.*).

The results demonstrated that the pre-trained model, without any fine-tuning, achieved performance comparable to its fine-tuned counterpart. This was largely

due to the effectiveness of the gradient-based thresholding method GradDiff, which showed improved robustness and consistency compared to conventional intensity-based thresholding. In addition, these findings suggest that key characteristics of anatomy and lesion may be visually preserved across age groups, and that the proposed framework can leverage this similarity through its encoding mechanism.

Two fine-tuning strategies were explored. Classical fine-tuning using a limited subset of paediatric cases proved more effective than the encoder-based strategy, which had previously shown strong results in generative tasks such as synthesis and SR. The reduced benefit of encoder adaptation may be explained by the similarity between source and target tasks, suggesting that domain-specific fine-tuning with this strategy did not prove beneficial in this setting.

In addition to the state-of-the-art publicly available dataset, the evaluation was conducted on a curated, multi-institutional cohort of paediatric cases, enabling assessment under realistic clinical conditions. These results demonstrate that the proposed weakly-supervised LDM framework can generalise beyond the tumour types represented in the brain tumor segmentation (BraTS) dataset, producing accurate predictions even for ependymoma and diffuse midline glioma (DMG). The approach also handles cases with incomplete brain coverage, as the encoding mechanism preserves existing anatomy without introducing spurious artefacts and restricting edits to relevant regions. At the same time, the analysis highlights a key limitation: reliance on complete multi-sequence inputs and curated preprocessing pipelines reduces flexibility when confronted with the variability of real-world clinical data. Even with this limitation, the generated healthy counterfactuals approach the performance of the supervised baseline in several cases, which itself demonstrates limitations in more challenging scenarios. Overall, these findings underline both the potential and the current constraints of the framework, showing that while it can generalise to heterogeneous clinical data and remain robust to moderate domain shifts, further refinement and validation are required for reliable clinical application.

6.2 Future work

While this research project has made substantial progress in advancing the field of weakly-supervised segmentation of paediatric brain tumours with DDPMs, several avenues for future work remain. This includes subsequent research that can build upon the findings and methodologies established in this thesis.

Managing interdependencies and optimising modularity

The modular nature of the proposed framework allows flexible adaptation to a variety of tasks, including lesion detection, lesion synthesis, and SR. This flexibility enables the composition of specialised components for each subtask, supporting the construction of tailored pipelines. However, this modularity also introduces vulnerability, particularly in the case of LDMs. The interaction between independently trained modules adds layers of complexity that are difficult to predict and interpret. While the reconstruction error of the first-stage model has been identified as a major influence on latent space quality, it is unlikely to be the only factor affecting downstream performance. Each module may respond differently depending on its objective, and their interplay can introduce confounding effects. The choice of hyperparameters further compounds this problem: selecting optimal values for guidance strength, sampling steps, encoding mechanisms, or sampling strategies proved highly sensitive to the specific approach and even to individual cases. The search space is vast, and current practice often relies on prior experience rather than systematic exploration. A more principled approach to parameter selection is therefore needed to ensure reliable outcomes, which can be inspired by the automatic adjustment strategies employed in frameworks like nnU-Net (Isensee et al., 2021).

Secondly, the dependency between modules implies that modifications to one component often necessitate retraining others, which can be time-consuming and resource-intensive. Although this is not necessarily more sensitive than training a single large-scale model, it adds a practical burden that must be carefully managed. Changes introduced at earlier stages may propagate through the pipeline, requir-

ing systematic evaluation to ensure stability. One promising strategy to alleviate this burden is the use of pretrained priors, as explored through the encoder-based conditioning mechanism in Chapter 4. By training a task-agnostic encoder, the framework can leverage already learned feature representations without re-optimising large backbone models. This approach reduces computational overhead, facilitates modular reuse, and improves scalability by isolating components that can generalise across tasks or domains. As such, it represents a viable step toward more efficient and maintainable weakly-supervised anomaly detection frameworks based on DDPMs.

In conclusion, future work should focus on developing structured approaches to analyse and manage the interdependencies between modules, ensuring that the benefits of modularity are preserved without compromising robustness. In particular, the influence of the first-stage model on latent space quality and downstream performance should be systematically investigated. Given that all subsequent components rely on its output, a clearer understanding of its role is essential for establishing reliable design choices and implementation guidelines.

Design of the anomaly detection approach

The anomaly detection method presented in this work is built around the generation of a healthy counterfactual, which is subsequently compared to the original image to localise potential lesions. While this counterfactual-based strategy enables weakly-supervised anomaly detection without pixel-wise labels, its probabilistic nature renders it sensitive to outliers and stochastic variations during sampling. These variations can propagate into the anomaly map and directly affect the accuracy of lesion detection. This work addressed part of this vulnerability through a gradient-based refinement strategy, GradDiff, which in ablation studies showed improved sensitivity and specificity for various lesion sizes when carefully tuned to the dataset at hand. It further demonstrated that computational demands may be reduced by disabling the memory-intensive latent gradient calculation and relying on classifier-free pseudo-gradients, with negligible loss in accuracy. Nonetheless, the method remains imperfect, and further developments are required to improve robustness,

particularly in edge cases or subtle lesions where perceptual discrepancies are minimal. The gradient-based refinement represents one component of this puzzle, as it improved detection performance but still requires further investigation to determine how its parameters and components can be balanced most effectively.

One promising direction lies in the integration of feature-space comparisons drawn from intermediate layers of the model, which may reveal more abstract discrepancies between real and counterfactual representations that are not perceptible in the image domain. Additionally, modality-aware weighting strategies could be explored to prioritise sequences with higher lesion visibility, such as T_2-fluid attenuated inversion recovery (T_2-FLAIR) for oedema or T_1 contrast-enhanced (T_1ce) for enhancing tumour cores. Finally, ensemble-based strategies or Bayesian extensions to the diffusion model could help quantify uncertainty, thereby allowing more cautious interpretation in low-confidence regions.

Fine-grained synthetic generation

While simple binary brain masks were sufficient to guide the generative process in this work, future advances in generative modelling should employ more anatomically and physically informed conditioning strategies. Anatomy-based guidance could specify the location of critical structures (e.g. ventricles, cortical grey matter, white matter, deep nuclei) to improve spatial plausibility and reduce anatomical artefacts (Bhattacharya et al., 2025). Physics-inspired approaches could explicitly model tissue displacement caused by tumour growth (Elazab et al., 2018, 2020; Hogea et al., 2007; Lipková et al., 2022). While current models learn this relation implicitly, incorporating biomechanical priors may yield more realistic deformations, particularly for small lesions where displacement is subtle yet clinically relevant. This would also enable tumour progression simulations, supporting treatment response assessment and intervention planning through predicted growth trajectories. Given the flexibility of the LDM framework, such conditioning could be integrated without major architectural changes, making it a promising direction for future research.

Furthermore, the integration of joint shape and appearance models represents a promising avenue for improving the detection of large-scale anomalies. Unlike current intensity-based methods, joint models can explicitly distinguish between pathological tissue appearance and the secondary morphological deformations (mass effect) caused by tumour growth (Bauer et al., 2012). Incorporating statistical shape priors or biomechanical growth simulations into the generative process would allow the framework to better accommodate the high degree of anatomical distortion seen in advanced cases. Recent evidence indicates that explicitly accounting for such deformations can improve model calibration and segmentation accuracy by up to 10% (Subramanian et al., 2023). This would be particularly valuable for ensuring that the generative "healthy" counterfactual accurately represents the displaced anatomy rather than merely synthesising healthy tissue within distorted structures.

Additionally, conditioning strategies may benefit from richer lesion representations that go beyond binary spatial masks. Incorporating probabilistic or latent embeddings of lesion characteristics could help preserve subtle pathological features through downsampling and improve guidance fidelity during sampling. A promising direction is the construction of synthetic paired datasets, where lesions are superimposed onto healthy brain images using generative models. Such datasets would provide explicit healthy-diseased pairs with exact spatial correspondence. They could be used to train SR models to consistently reconstruct the healthy high-resolution (HR) version of an input image, irrespective of whether the low-resolution (LR) input contains a lesion. This could unify the tasks of SR and healthy counterfactual generation attempted in Section 4.4.5, reducing inference time and improving consistency.

Generalisability and robustness

The proposed framework demonstrates that weakly-supervised anomaly detection substantially reduces reliance on dense annotations while maintaining high segmentation performance. This lower level of supervision eases scalable training and deployment, especially in clinical settings where manual annotation is often infeasible. Importantly, this lays a foundation for models that generalise beyond constrained

datasets to accommodate broader clinical variability. Initial steps were taken in Chapter 5, where the framework was applied to the data-scarce paediatric domain, demonstrating strong robustness and generalisability.

One key challenge in real-world deployment is the heterogeneity of clinical data. Unlike curated datasets, clinical images vary in resolution, acquisition protocols, scanner hardware, and patient demographics. Moreover, anatomical variation adds another layer of complexity. Patients with congenital abnormalities, prior surgical resections, or trauma may exhibit highly irregular brain structures that do not correspond to either healthy or classically pathological appearances. Models must therefore not only detect anomalies, but also discern when such deviations are clinically meaningful. Improving robustness under these conditions is central to ensuring that weakly-supervised approaches retain utility in diverse clinical scenarios.

To accommodate such clinical variability, more flexible solutions are needed that go beyond rigid modelling assumptions and can adapt to heterogeneous data. DDPMs, with their strong capacity to model complex distributions and their susceptibility to conditioning, are particularly well suited for this task. Potential applications include to synthesise missing complimentary MRI sequences, simulate different scanner strengths (e.g., 1.5 T to 3 T conversions), or perform image-to-image translation under varying acquisition protocols. These capabilities would be especially valuable in clinical environments with incomplete imaging or heterogeneous scanner infrastructure, helping to harmonise data and provide a more consistent basis for analysis. By acting as a tool for image harmonisation, DDPMs can standardise intensity distributions and contrast profiles across different scanner vendors and field strengths, providing a consistent anatomical basis for downstream analysis and mitigating the "site-effects" that often degrade model performance in multi-centre studies.

To this end, expanding the semantic resolution of weakly-supervised anomaly detection with DDPMs is a promising direction. Beyond disease detection, models could benefit from incorporating auxiliary subject-level labels such as tumour subtype, grade, location, or even genetic and molecular markers. Complementary information from radiology reports, pathology findings, or genomics could further contextualise

the imaging signal, allowing the model to disambiguate between phenotypically similar but clinically distinct entities. Such auxiliary signals could also help identify distribution shifts that cause the anomaly detection framework to segment some cases accurately while failing in others. This richer supervision could enhance the specificity of predictions and provide a more holistic view of each individual.

Another opportunity lies in using this weakly-supervised framework as a base for pre-training large-scale image DDPMs. In natural language processing, unsupervised pre-training on unstructured data has yielded powerful general-purpose representations. By analogy, DDPMs trained on large, weakly-annotated imaging datasets may learn transferable features that capture the diversity of real clinical populations. Because the approach relaxes the requirement for dense label information, it is inherently more scalable and allows access to substantially larger datasets for pre-training, further strengthening the robustness of the learned representations. These pretrained models could then be adapted for specific tasks through targeted fine-tuning with small amounts of labelled data, e.g. using the encoder-based approach outlined in this research project. The success of this approach depends critically on the quality of the encoding mechanism, which must produce intermediate representations that preserve anatomical and pathological relevance.

Finally, the choice and scope of pre-training data will strongly influence the generalisation capacity of these models. To be clinically reliable, pretrained models must be exposed to the full spectrum of anatomical variants, disease presentations, and acquisition artefacts encountered in routine imaging. This includes structurally abnormal yet non-pathological cases, which are often underrepresented in curated datasets. Moreover, future work should investigate the applicability of the proposed framework to other lesion types, such as haemorrhagic abnormalities, multiple sclerosis plaques, or metastatic foci. Demonstrating robustness across these diverse pathologies would further validate the generalisability of the approach and provide a unified foundation for weakly-supervised lesion detection across the central nervous system. Ensuring resilience to this broad clinical spectrum is key to building models that are not only accurate, but also robust and trustworthy in real-world deployment.

Clinical integration

The ultimate goal of weakly-supervised segmentation is to provide a clinically useful tool that can assist radiologists in diagnosing and monitoring brain tumours. However, integrating these models into clinical workflows presents several challenges. First, the interpretability of model predictions must be improved to ensure that clinicians can trust and understand the rationale behind the segmentation results. This is particularly important in high-stakes environments like oncology, where treatment decisions are based on imaging findings. The anomaly detection framework developed in this thesis addresses this challenge by generating healthy counterfactuals of pathological inputs. Beyond identifying the location of a lesion, this approach also reveals the underlying anatomy that would be present in the absence of disease. Such information could provide additional context to the assessing physician, supporting more informed interpretation of imaging findings and potentially aiding clinical decision-making.

Second, the performance of weakly-supervised models must be validated against established benchmarks and clinical standards. This includes comparing model outputs with expert annotations and assessing their impact on diagnostic accuracy and treatment planning. Rigorous validation studies are essential to demonstrate that these models can reliably augment human expertise rather than replace it.

Another important avenue concerns the role of hospital infrastructure in supporting data curation and validation (Willemink et al., 2020). With weakly-supervised brain tumour segmentation, the reliance on fine-grained annotations is reduced, allowing datasets to be curated more rapidly and at larger scale. Streamlining data pipelines within hospitals to prepare imaging data directly after acquisition (while adhering to ethical guidelines and anonymisation protocols) would increase the availability of training and validation material. Such infrastructure could further integrate complementary clinical markers beyond imaging, creating richer multimodal datasets with impact that extends beyond segmentation alone. Establishing these pipelines would not only accelerate model development but also enhance the robustness and

clinical relevance of weakly-supervised approaches. Furthermore, federated learning offers an important opportunity to scale these models across institutions without the need for raw data exchange. In paediatric oncology, where datasets are inherently scarce and highly sensitive, federated learning would allow for the collaborative training of anomaly detection models on decentralised hospital servers. This approach preserves patient confidentiality while ensuring that the resulting models are exposed to a sufficiently diverse range of pathological variants and acquisition protocols to be truly robust in a real-world clinical setting.

Beyond infrastructure, the availability and quality of clinical data remain a central limitation. This work highlights the dependence of current approaches on the BraTS dataset and its co-registration pipeline, which presumes the availability of all complementary MRI sequences. In clinical practice, such complete data are often unavailable, which restricts the applicability of models trained under these assumptions. Addressing this constraint requires more flexible strategies for handling incomplete or heterogeneous input. One option is to use generative models to synthesise missing sequences, thereby conforming to established processing standards. Another direction is to move beyond the rigid multi-sequence paradigm altogether. For instance, modality-specific encoder layers within the LDM could be trained on datasets with complete coverage and occasionally masked during training. This approach could reduce the reliance on every MRI sequences and may permit the model to operate with any subset of available inputs. Such a design could also accommodate non-standard modalities such as perfusion- and diffusion-weighted imaging or functional MRI, which are rarely available but clinically valuable. Achieving this flexibility would require adaptations to the preprocessing and co-registration pipelines, but it represents an important step toward clinical integration without excluding individuals based on data availability.

Furthermore, user-friendly interfaces and tools must be developed to facilitate the adoption of these models in clinical practice. This includes integrating segmentation results into existing radiology software, providing visualisations that highlight model confidence, and enabling easy interaction with model outputs.

Aside from technical and clinical hurdles, the deployment of generative models in healthcare must address considerable data privacy risks, particularly regarding model memorisation. Recent studies have demonstrated that large scale diffusion models can inadvertently memorise specific training samples, potentially exposing sensitive patient data (Carlini et al., 2021). This risk is particularly acute in medical imaging, where recent findings on 3D latent diffusion models suggest that up to 37% of patient data can be detected as memorised under certain training configurations (Dar et al., 2025). Such vulnerabilities can be exploited through membership inference attacks, which determine if a patient's data was used in training, or model inversion attacks, which attempt to reconstruct images from the model's parameters (Matsumoto et al., 2023; Packhäuser et al., 2022). In paediatric oncology, where datasets are small and anatomical features are unique, these risks are substantial. To mitigate these threats, future iterations of this framework should incorporate practical safeguards such as differential privacy to provide mathematical guarantees against data leakage (Dockhorn et al., 2023). Additionally, robust anonymisation pipelines and post-generation copy detection mechanisms are essential to ensure that synthetic data remains truly private and non-recoverable (Kaissis et al., 2021).

6.3 Final remarks

While accurate lesion segmentation may not yet directly translate into specific therapeutic interventions for all brain tumour cases, the ability to reliably detect and delineate anomalous regions represents a pivotal advancement, particularly in paediatrics. This work has addressed key gaps in weakly-supervised segmentation, improving sensitivity to small lesions, enhancing spatial context through volumetric modelling, and advancing generalisability across diverse clinical populations. These capabilities lay the groundwork for robust, scalable diagnostic tools that support clinicians in forming a comprehensive view of brain pathology. As precision medicine, surgical planning, and genomic analysis continue to evolve, they will increasingly depend on accurate, automated imaging assessments. Prioritising early and precise

detection is therefore crucial for current clinical care and forms the foundation for future therapeutic progress.

The guiding question posed in Section 2.2.2

> ***Can we build models that no longer depend on fine-grained, expert-level annotations?***

can now be answered with cautious optimism. Relying solely on the creation of ever-larger, fully standardised multi-centre datasets is unlikely to provide a sustainable solution, given the costs and coordination challenges involved. Progress instead depends on methods that can extract value from heterogeneous and incomplete clinical data. The weakly-supervised anomaly detection framework developed in this thesis provides such a pathway: it reduces reliance on dense annotations, narrows the performance gap to fully supervised models, and has shown potential to detect anomalies beyond the training distribution. These qualities highlight its scalability across datasets and its capacity to operate with reduced dependence on costly expert annotations. Together, these advances move artificial intelligence closer to realising its full potential and support a future where early diagnosis and intervention are more accessible, consistent, and effective across all patient populations.

Appendix A

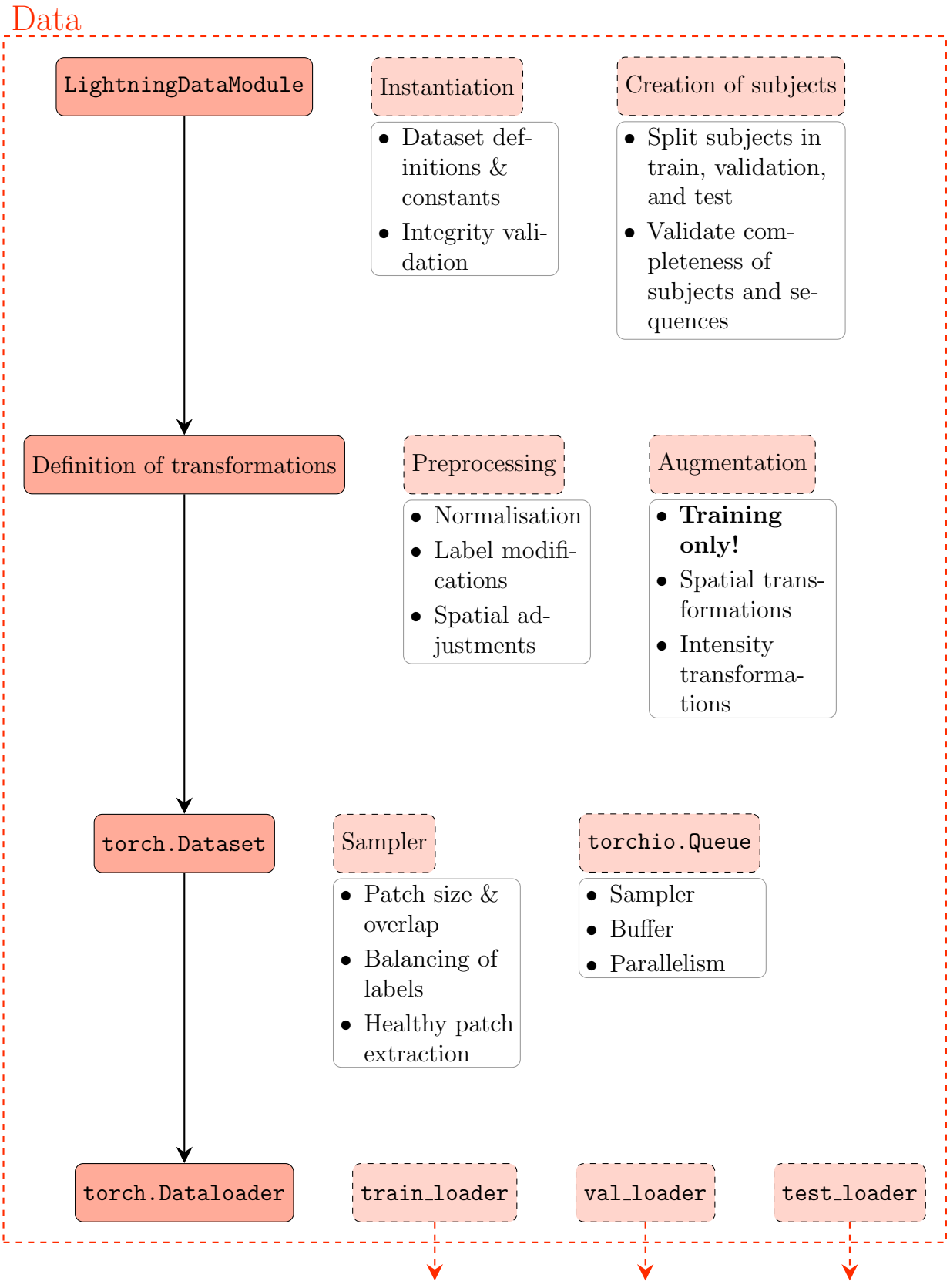
3D latent diffusion model

A.1 Conceptualisation of model framework

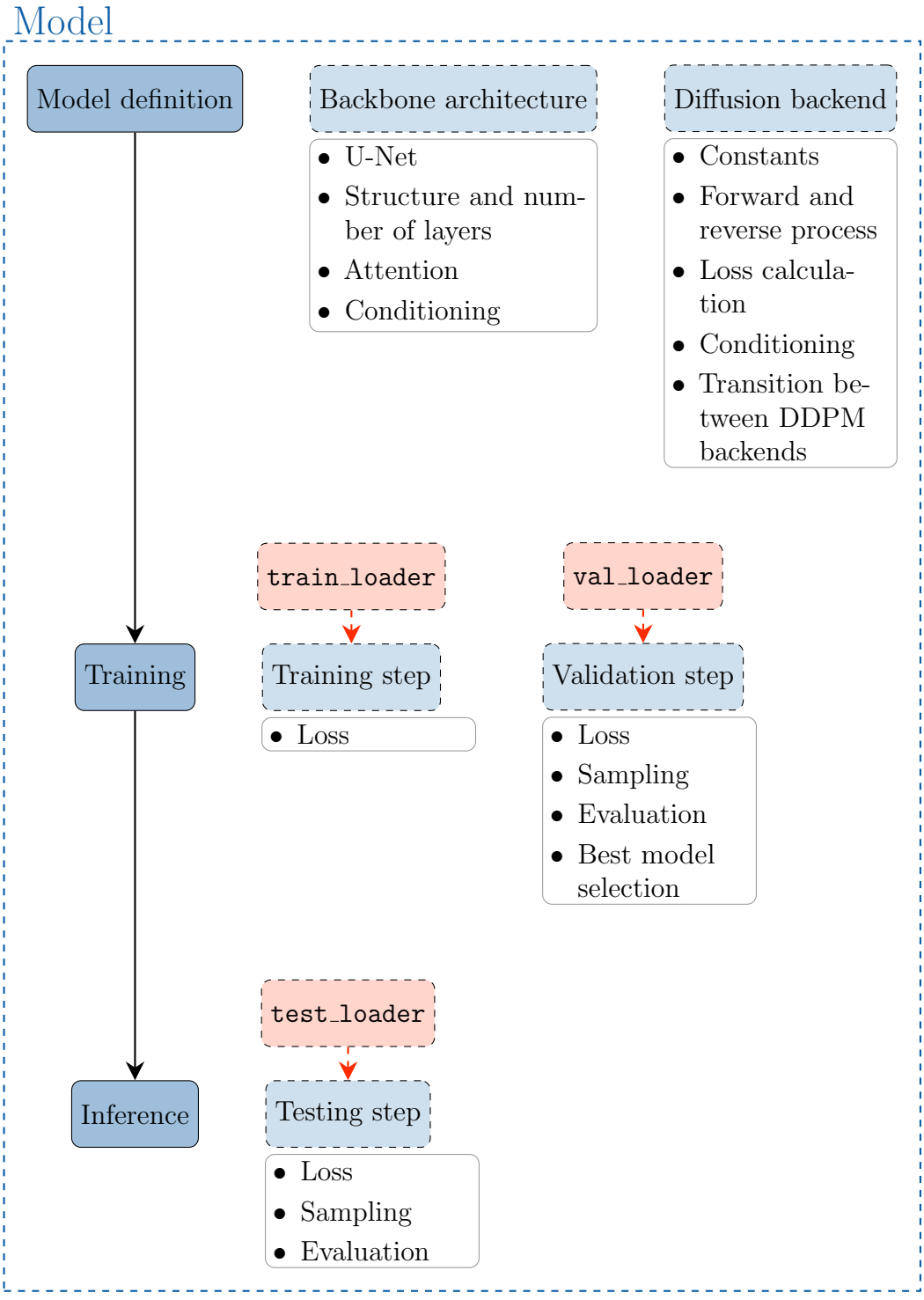
The design of the model framework is based on the principles of modularity and separation of concerns. `pytorch-lightning` (Falcon & The PyTorch Lightning team, 2019) serves as the overarching framework, chosen for its modular and extensible design to efficiently incorporate all aforementioned requirements. This framework enables rapid prototyping and simplifies the integration of advanced features such as multi-GPU training, checkpointing, and logging, which streamline the tracking and management of the training process. Fundamentally, it abstracts both data and model components into separate modules, facilitating the development of a highly adaptable and scalable framework. `pytorch-lightning` fundamentally builds on `PyTorch` (Paszke et al., 2019), which offers a flexible and efficient backend for Python-based implementation of neural network (NN) operations, including data processing, model design and training.

A.1.1 Data processing pipeline

The ***data processing pipeline*** is specifically designed to address the unique requirements of medical imaging data, ensuring efficient loading, preprocessing, and augmentation. Building on the specialised medical imaging library `torchio` (Pérez-



(a) Data processing pipeline



(b) Model architecture and training pipeline

Figure A.1: Detailed schematic of the data processing and model training pipelines, extending the overview in Fig. 3.1. (a) illustrates dataset composition, applied preprocessing transformations, and integration with `torch` and `pytorch_lightning`. (b) details the definition of the model, its components, and the subsequent training and evaluation procedures.

García et al., 2021), the pipeline provides advanced data handling and augmentation capabilities specifically tailored for medical data.

The pipeline is built around an adapted version of the `LightningDataModule` from `pytorch-lightning`, a core component of its data and model abstraction framework that governs data loading and preprocessing workflows. The first step involves loading and validating the dataset definition file, which contains general information about the dataset and the paths to all available images or volumes; the paths are grouped by patient in the case of the BraTS dataset. Each sequence is typically stored in the Digital Imaging and Communications in Medicine (DICOM) format or, more commonly in research settings, in the Neuroimaging Informatics Technology Initiative (NIfTI) format, which offers a more compact and efficient representation of imaging data. The NIfTI format not only reduces file size but also facilitates seamless integration with computational tools and platforms, making it a preferred choice for neuroimaging research (Larobina & Murino, 2014).

With a complete dataset stored locally, the next step is to create `torchio.Subject` instances for each patient. A subject represents a single patient or examination and is designed to consist of multiple connected MRI sequences (see Section 2.1.1). Based on user-specified splits for training, validation, and testing, the annotated subjects are subsequently divided into their respective subsets.

The subsequent stage defines the transformations applied to each subject within the specified split. These transformations encompass standard preprocessing operations, such as normalisation, spatial adjustments, and label modifications, as well as augmentation techniques aimed at enhancing data diversity in light of the limited sample size. The augmentations are categorised into spatial and intensity transformations, with the former modifying the spatial properties of the data, such as rotations, scaling, and elastic deformations, while the latter adjusts the intensity values of the images, including contrast and brightness modifications. An optional post-processing step may also be applied, typically involving a spatial transformation to adjust the appearance for logging, due to the differing orientation standards between the backbone DICOM and NIfTI data loaders in `torchio`.

Importantly, augmentations are applied exclusively to the training split to increase the variability and volume of training data, while pre- and post-processing steps are applied to all subject splits. Additionally, transformations are not applied immediately; instead, they are deferred until the data is loaded through the `torch.Dataloader`, thereby offloading the computational burden to the moment when the data is needed for training.

Once the transformations are defined, a `torch.Dataset` instance is created for each split, which is necessary for subsequent data loading via the `torch.Dataloader`. The `torch.Dataset` is implemented using `torchio.Queue`, enabling efficient parallel preprocessing of data while the model is being trained. This approach ensures that data is preprocessed and available when required, optimising workflow and minimising potential bottlenecks. A crucial element of the queue is the sampler, which extracts patches of a predefined size from each subject. This sampler is essential for obtaining healthy patches, which are then used to generate healthy counterparts during the diffusion process. By adjusting the probability of selecting both diseased and healthy patches based on the ground-truth label mask, the sampler ensures unbiased model training. Additionally, the `torchio.Queue` acts as a buffer for the sampled patches, which are shuffled once the queue reaches a specified length. This facilitates stochastic patch-based training, ensuring each batch contains patches from multiple subjects, thereby increasing data variability within each iteration.

Finally, the `torchio.Queue` is passed to the `torch.Dataloader`, which handles efficient data loading parallel to model training. The `torch.Dataloader` is responsible for batching the data, distributing the workload across multiple workers to maximise efficiency, and prefetching data to minimise idle time during training. Serving as the final component of the data pipeline, the `torch.Dataloader` ensures a continuous stream of preprocessed data for training, validation, and testing.

The modular design of the pipeline, with its adaptable transformations and sampler, enables customisation to meet the specific requirements of different datasets and problem scenarios. This flexibility ensures compatibility with a wide range of tasks. By adhering to a standardised interface and dataset definitions, datasets can

be seamlessly interchanged, facilitating streamlined integration and efficient data preparation.

A.1.2 Backbone model architecture and diffusion backend

The design of the ***model architecture*** (shown in Figure A.1b) is centred around two primary components: the ***backbone model architecture*** and the ***diffusion backend***. These components are integrated into the `LightningModule` from `pytorch-lightning`, forming the second aspect of their separation of data and model. The backbone model is tasked with learning and predicting the reverse process described in Equation (2.18). To ensure that the output dimensions align with the input, the backbone follows a U-Net structure consisting of an encoder, bottleneck, and decoder. Its modular design incorporates key elements such as convolution and attention, which are organised into residual blocks and stacked across layers to extract the rich features necessary for accurate predictions of the mean $\mu_\theta(\boldsymbol{x}_t, t)$ and, optionally, the variance $\Sigma_\theta(\boldsymbol{x}_t, t)$ of the forward process. Additionally, the backbone model includes specialised feature extractors for the timestep t and class embedding C, which facilitate conditioning of the diffusion process based on temporal and class-specific information. The number of available training parameters is influenced by several hyperparameters, including, but not limited to:

- ***Hidden channels***: output feature channels of the first convolutional layer preceding the encoder
- ***Channel factor***: depth of the encoder (and decoder) and specifies the multiplicative factor given the hidden channels for each layer
- ***Number of residual blocks***: number of residual blocks per layer comprised of convolutional and attention layers
- ***Attention resolution***: spatial resolution at which an attention layer is applied

As the model is defined as a `LightningModule`, it defines the forward pass required for loss calculation, and individual steps for training, validation and testing.

The diffusion backend defines the forward process of the diffusion model and includes methods for the backward process in DDPM and its variants (e.g., denoising

diffusion implicit model (DDIM)). It provides essential functionality for efficient loss computation and conditional sampling, particularly for generating healthy counterfactuals, a key focus of this research project. Serving primarily as an interface to the diffusion process, the backend abstracts its complexity and ensures seamless integration with the backbone model, which is then passed to various functions for training and sampling. Similar to the other components, the diffusion backend is modular, compartmentalising the diffusion process into forward, backward, and generative steps. This structure allows for easy extension and adaptation to various diffusion models and tasks.

A.2 Medical LDM framework

Table A.1: Detailed results of the first-stage models for L_1 reconstruction error.

PE	Codebook	MRI	L1 Max	L1 95%	L1 99%
None	A	T_1w	0.0649	0.0018	0.0048
		T_1ce	0.0618	0.0024	0.0053
		T_2w	0.0546	0.0024	0.0055
		T_2-FLAIR	0.0683	0.0022	0.0058
	B	T_1w	0.0648	0.0018	0.0050
		T_1ce	0.0637	0.0024	0.0053
		T_2w	0.0576	0.0025	0.0057
		T_2-FLAIR	0.0686	0.0023	0.0060
	C	T_1w	0.0678	0.0018	0.0049
		T_1ce	0.0660	0.0024	0.0053
		T_2w	0.0620	0.0024	0.0057
		T_2-FLAIR	0.0718	0.0021	0.0057
Sinusoidal	A	T_1w	0.0645	0.0018	0.0049
		T_1ce	0.0629	0.0024	0.0052
		T_2w	0.0568	0.0024	0.0054
		T_2-FLAIR	0.0704	0.0023	0.0059
	B	T_1w	0.0634	0.0018	0.0050
		T_1ce	0.0606	0.0024	0.0054
		T_2w	0.0544	0.0026	0.0058
		T_2-FLAIR	0.0672	0.0024	0.0061

Table A.2: Detailed results of the first-stage models for L_2 reconstruction error.

PE	Codebook	MRI	L2 Max	L2 95%	L2 99%
None	A	T_1w	0.0272	0.0000	0.0002
		T_1ce	0.0251	0.0001	0.0002
		T_2w	0.0203	0.0000	0.0002
		T_2-FLAIR	0.0324	0.0000	0.0002
	B	T_1w	0.0298	0.0000	0.0002
		T_1ce	0.0267	0.0001	0.0002
		T_2w	0.0265	0.0000	0.0002
		T_2-FLAIR	0.0311	0.0000	0.0002
	C	T_1w	0.0300	0.0000	0.0002
		T_1ce	0.0288	0.0001	0.0002
		T_2w	0.0287	0.0000	0.0002
		T_2-FLAIR	0.0323	0.0000	0.0002
Sinusoidal	A	T_1w	0.0253	0.0000	0.0002
		T_1ce	0.0267	0.0001	0.0002
		T_2w	0.0236	0.0000	0.0002
		T_2-FLAIR	0.0359	0.0000	0.0002
	B	T_1w	0.0286	0.0000	0.0002
		T_1ce	0.0250	0.0001	0.0002
		T_2w	0.0218	0.0000	0.0002
		T_2-FLAIR	0.0320	0.0000	0.0002

A.3 Experimental evaluation of the 3D-LDM

Table A.3: Comparison of specificity scores across baseline methods and the proposed 3D-LDM configurations. Baselines include 2D-CAM[a], 2D-WS-MTST[b], 3D-VQ-GAN, and 2D-DDPM. The proposed 3D-LDM is evaluated with both DDIM and EDICT sampling strategies.

	Model	Encoding	PE	C	N		
					50	100	150
Baseline	2D-CAM[a]	-	-	-		0.9994	
	2D-WS-MTST[b]	-	-	-		0.9994	
	3D-VQ-GAN	-	-	-		0.9557	
	2D-DDPM	DDIM	-	0	0.8970	0.8970	0.9015
				5	0.9722	0.9707	0.8218
				10	0.9767	0.9766	0.9026
Proposed	3D-LDM	DDIM	sin	0	0.9983	0.9963	0.9949
				5	0.9680	0.9583	0.9490
				10	0.9652	0.9581	0.9409
		EDICT	sin	0	0.9994	0.9974	0.9974
				5	0.9932	0.9927	0.9824
				10	0.9866	0.9833	0.9458

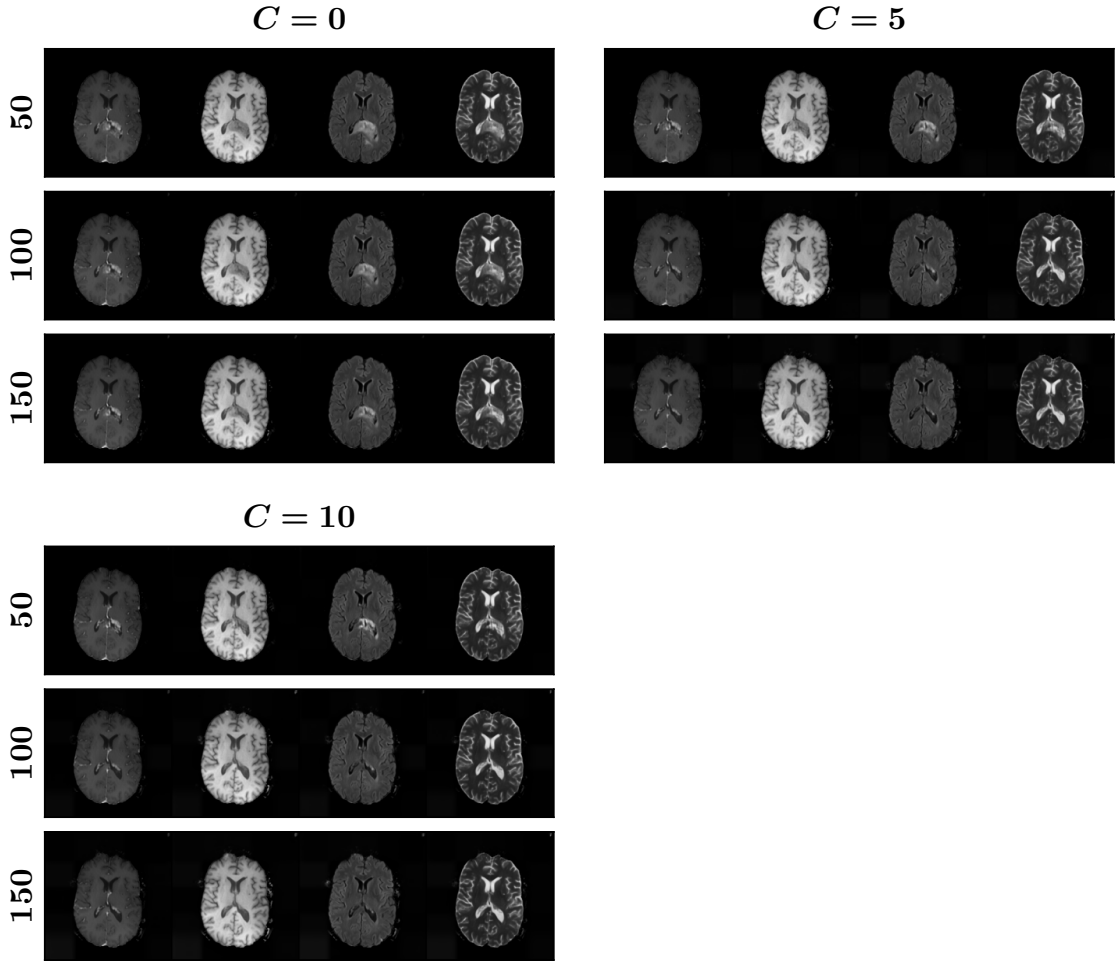
[a](Z. Chen et al., 2022)
[b](H. Chen et al., 2023)

Table A.4: Normalised inference times of the tested models in seconds per subject. Reported times are approximate due to differences in hardware configurations during inference.

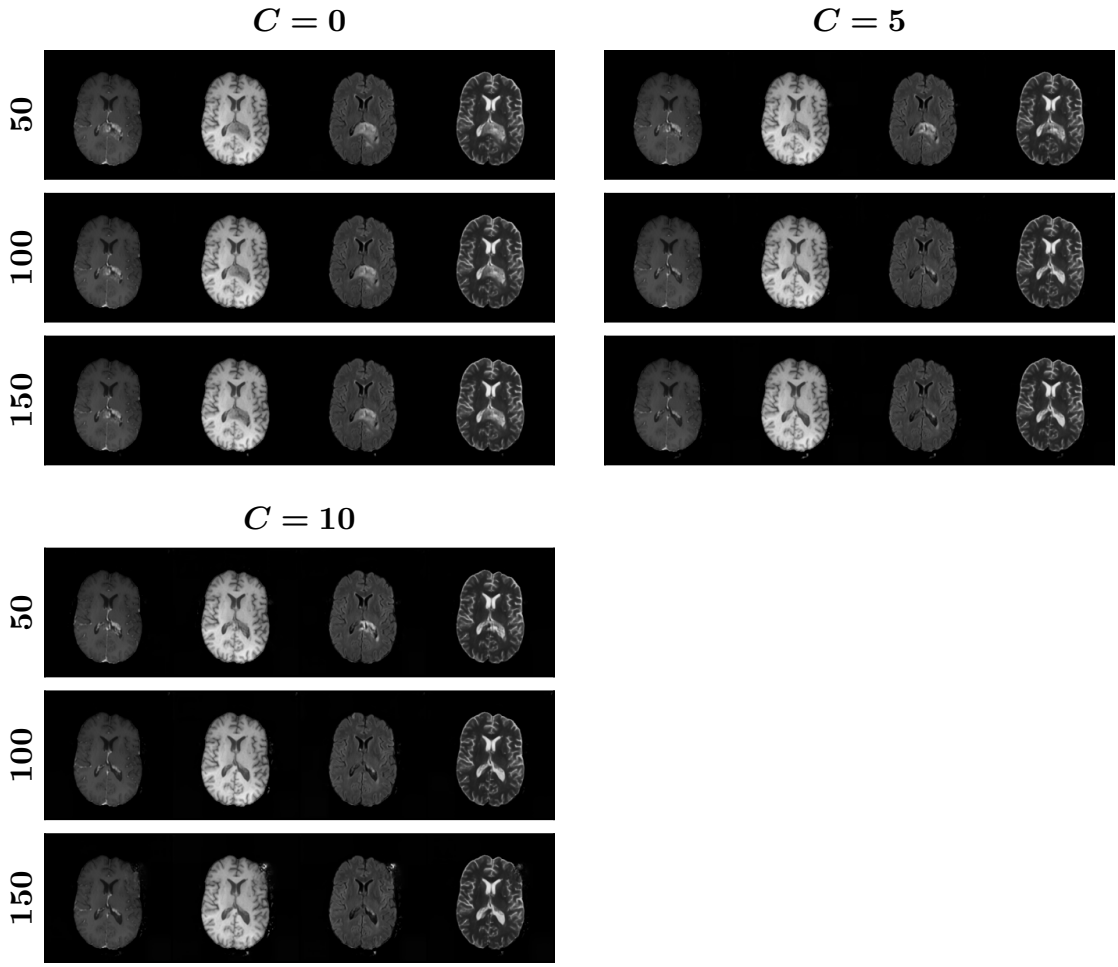
	Model	Encoding	PE	C	N		
					50	100	150
Baseline	2D-CAM [a]	-	-	-		0.3215	
	2D-WS-MTST [b]	-	-	-		0.3215	
	3D-VQ-GAN	-	-	-		12.7269	
	2D-DDPM	DDIM	-	0	235.7900	472.1925	483.1712
				5	353.4625	709.0225	711.9840
				10	353.4500	718.7200	711.9462
Proposed	3D-LDM	DDIM	sin	0	90.7653	159.8493	228.6920
				5	129.3413	229.5520	335.0627
				10	125.9213	230.6840	335.1173
	3D-LDM	EDICT	sin	0	160.1036	307.0292	454.1698
				5	305.5820	599.5135	892.4829
				10	305.3964	599.0560	893.9443

[a] (Z. Chen et al., 2022)

[b] (H. Chen et al., 2023)



(a) DDIM



(b) EDICT

Figure A.2: Healthy counterfactual generation for a single subject in a central brain slice, comparing DDIM and EDICT. Columns vary the classifier strength C, while rows vary the number of encoding steps N. The generated counterfactuals are masked with the input brain mask to suppress background artefacts, demonstrating the dependence of the generation process on C and N under the two sampling strategies.

Appendix B

Small lesion detection

B.1 Generating small brain tumours

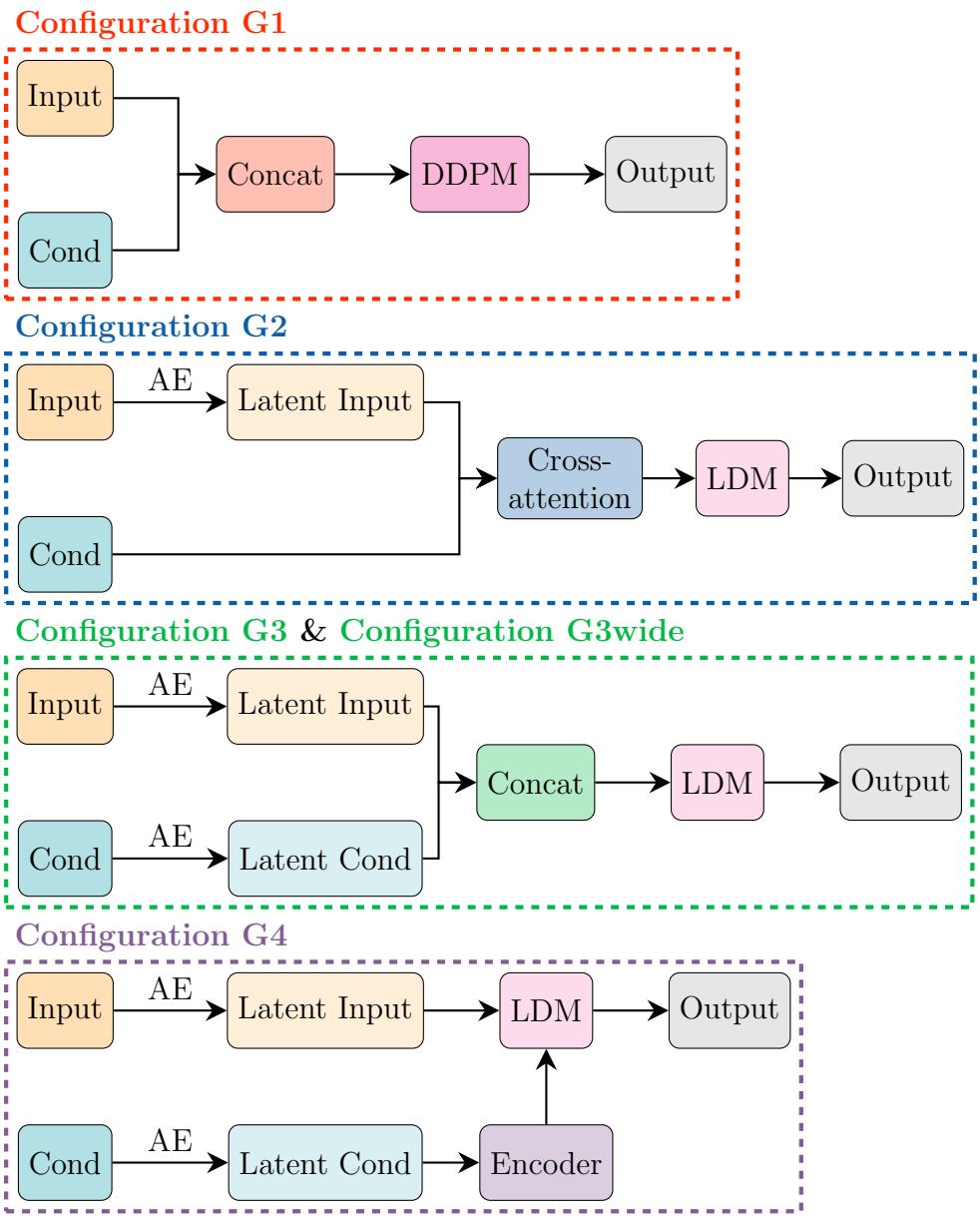
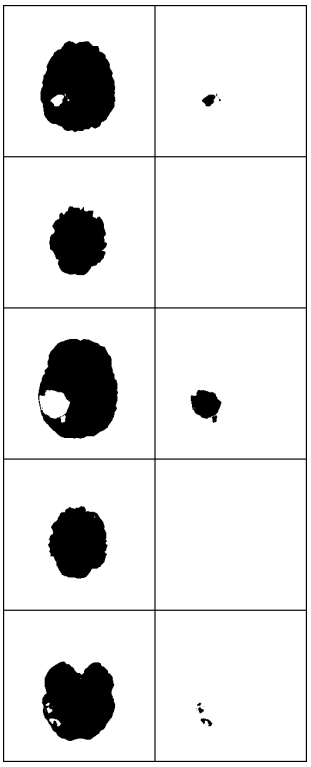
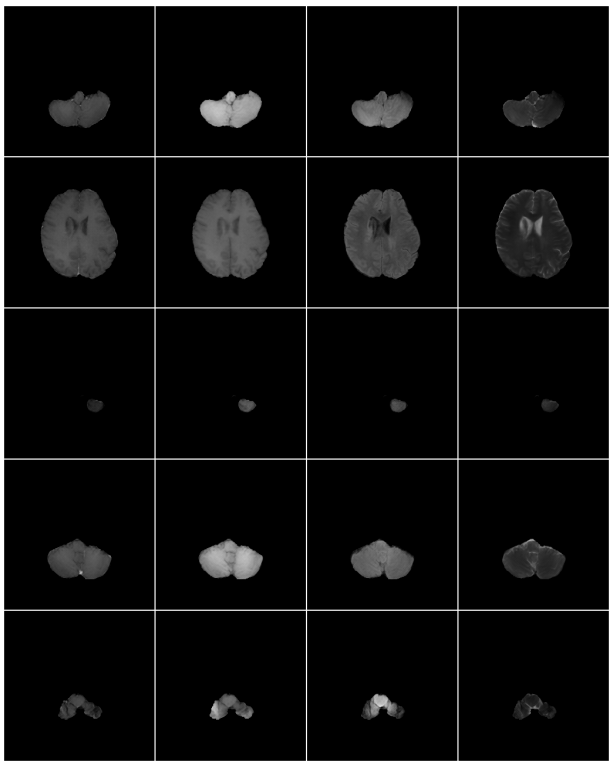


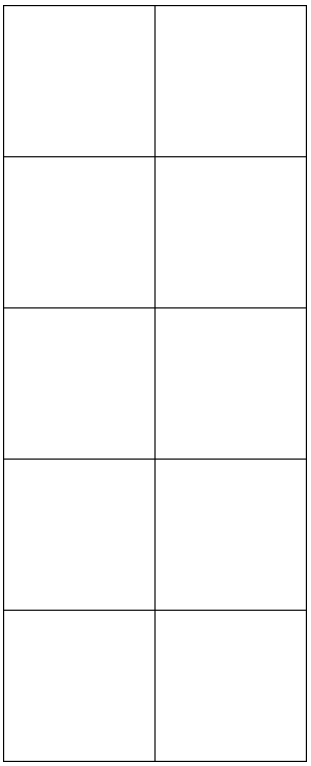
Figure B.1: Schematic overview of the conditioning mechanisms for small lesion generation, corresponding to the configurations in Table 4.1. Concat: concatenation; Cond: conditioning signal (binary lesion mask).



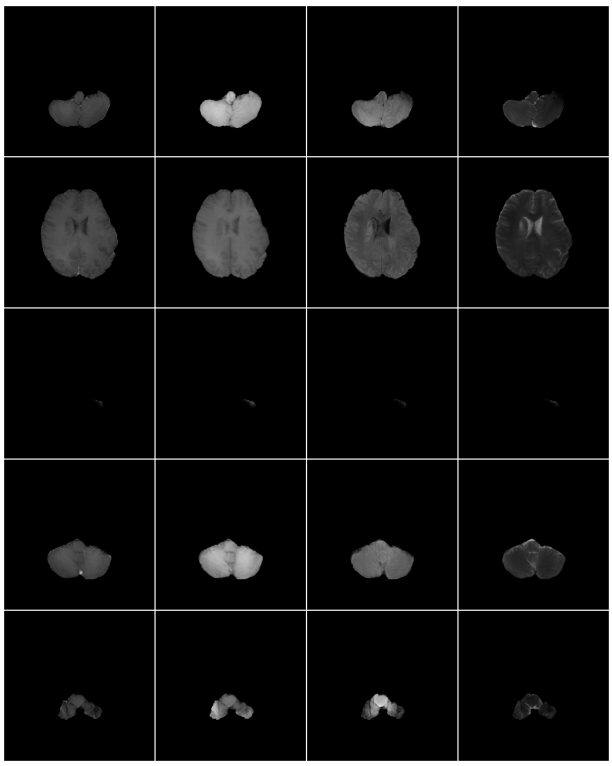
(a) Normal conditioning



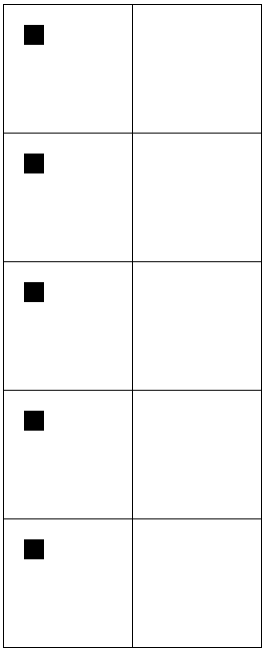
(b) Generated sample of normal conditioning



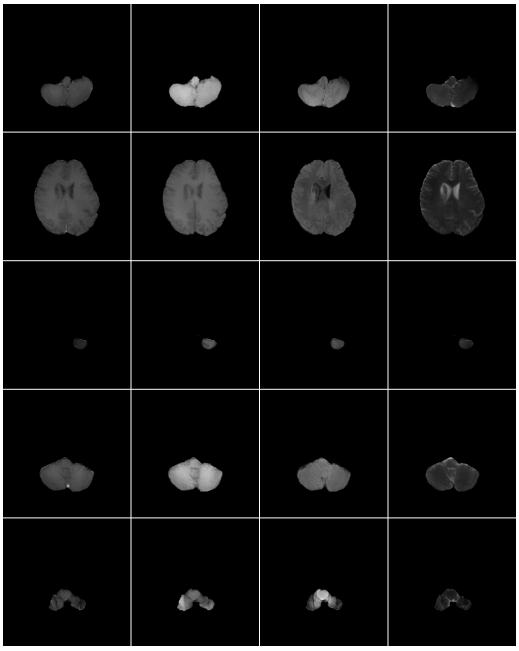
(c) Null conditioning



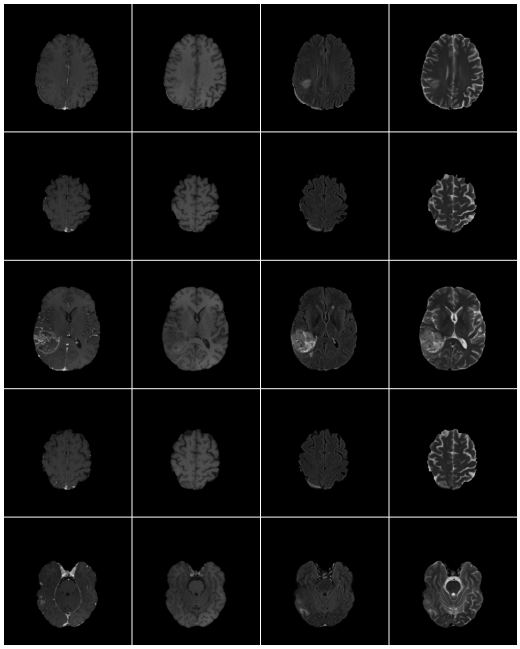
(d) Generated sample of null conditioning



(e) Synthetic conditioning

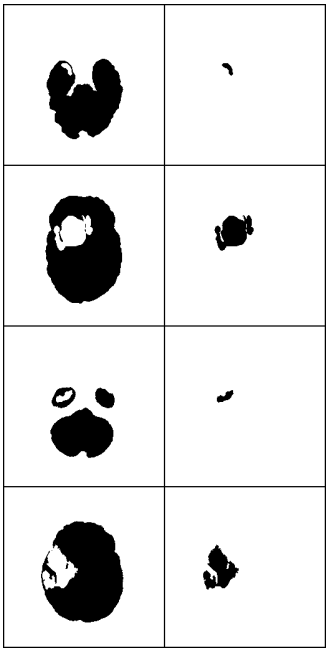


(f) Generated sample of synthetic conditioning

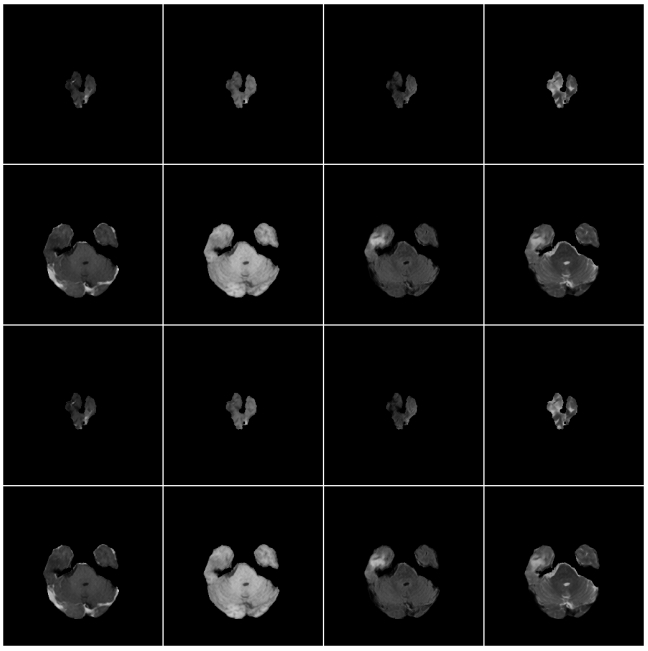


(g) Reference sample

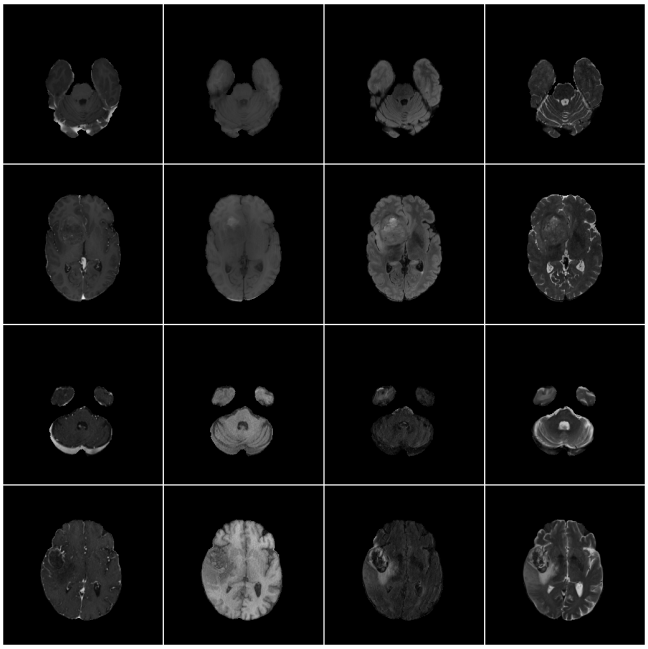
Figure B.2: Generated brain images using cross-attention conditioning for various conditioning signals. (c), (a) and (e) show the one-hot encoded brain mask (left column) and optional lesion target (right column) used as conditioning input. (d), (b) and (f) display the corresponding generated samples, whereas (g) shows the real MRI from which the conditioning signal was derived. Rows correspond to different 2D slices from individual subjects. Columns in (d), (b), (f) and (g) correspond to MRI sequences ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.



(a) Conditioning signal

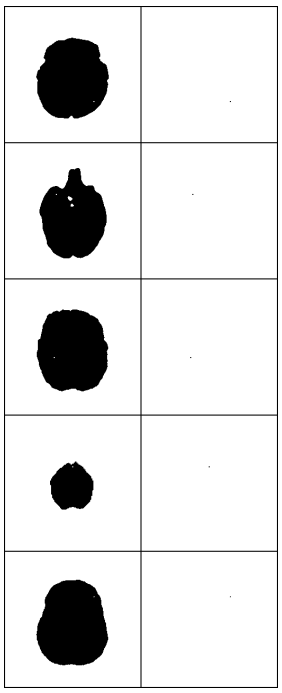


(b) Generated samples

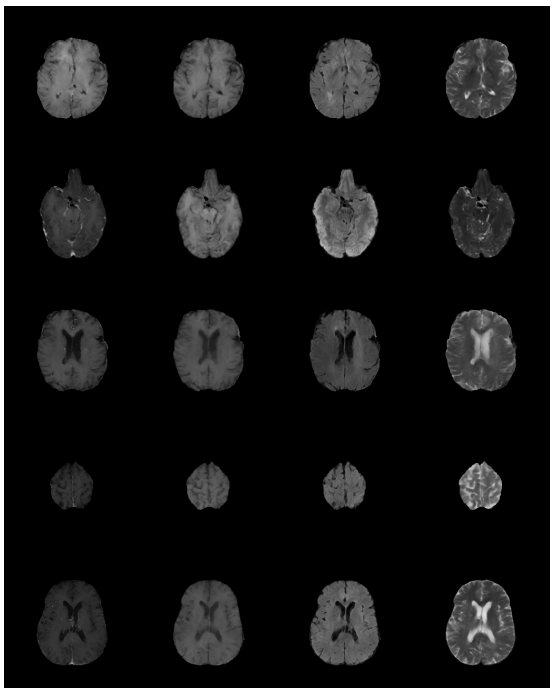


(c) Reference sample

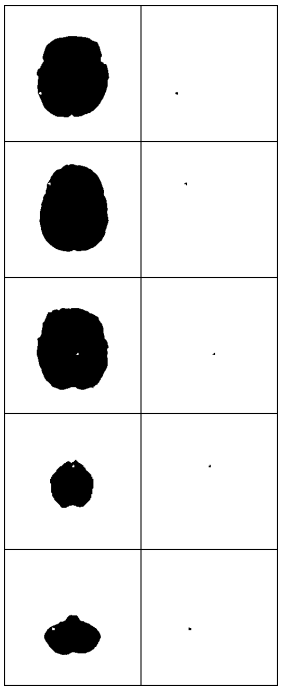
Figure B.3: Generated brain images using multi-stage cross-attention conditioning. (a) shows the one-hot encoded brain mask (left column) and optional lesion target (right column) used as conditioning input. (b) displays the corresponding generated samples, whereas (c) shows the real MRI from which the conditioning signal was derived. Rows correspond to different 2D slices from individual subjects. Columns in (b) and (c) correspond to MRI sequences ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.



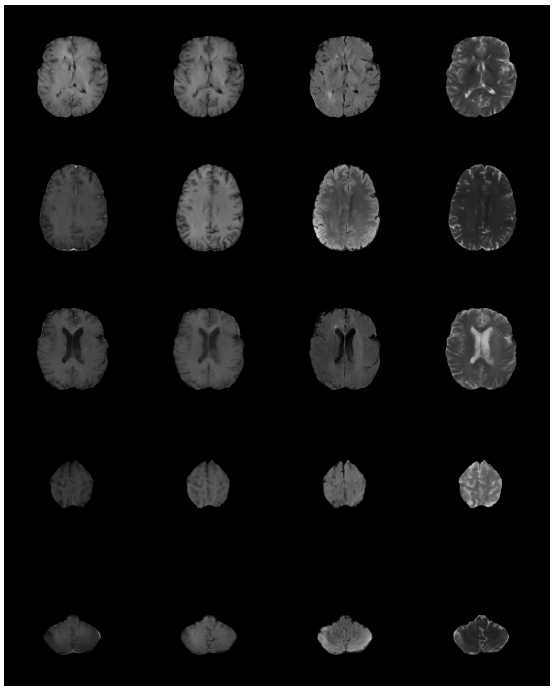
(a) Conditioning: 2 mm



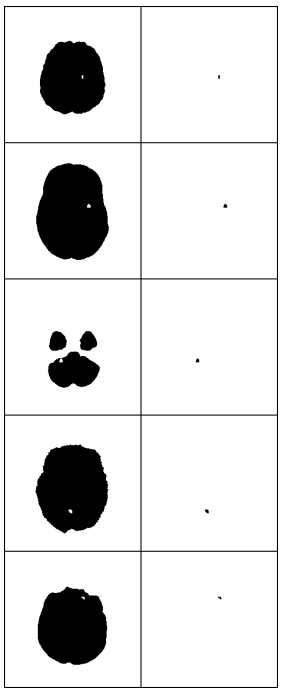
(b) Generated: 2 mm



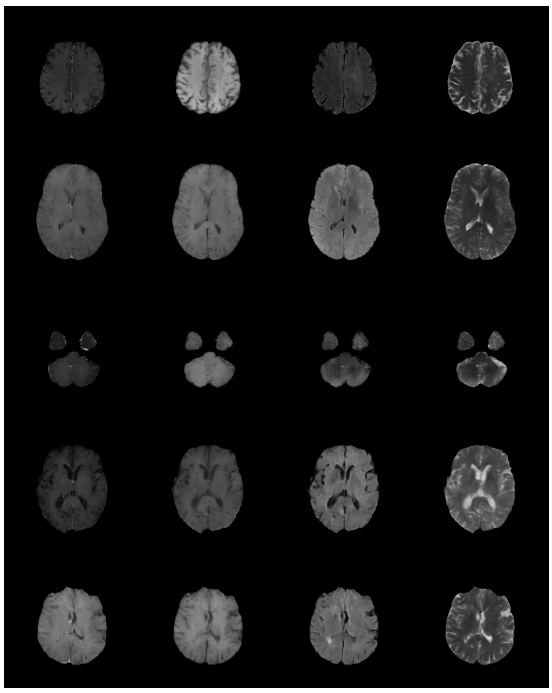
(c) Conditioning: 5 mm



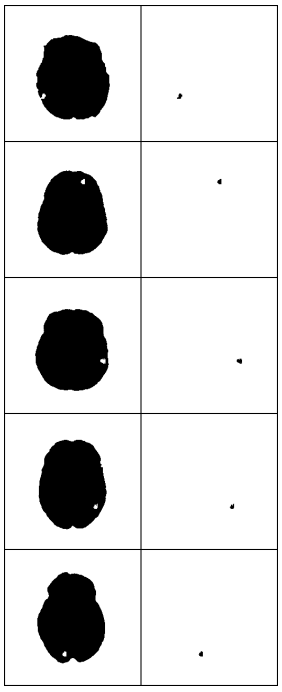
(d) Generated: 5 mm



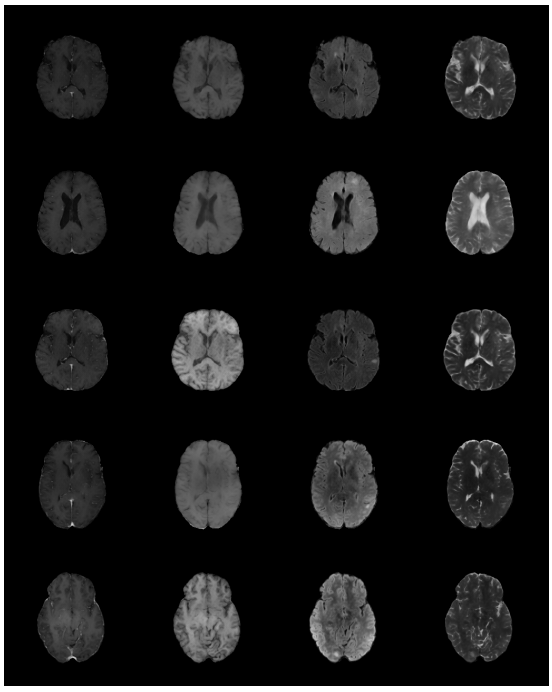
(e) Conditioning: 7 mm



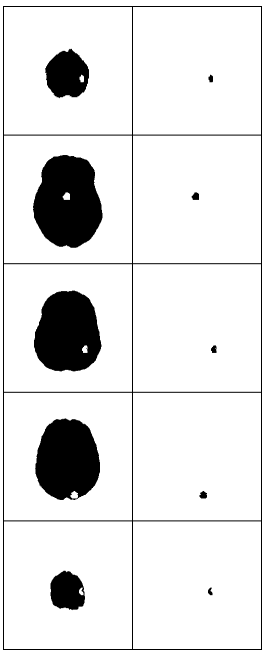
(f) Generated: 7 mm



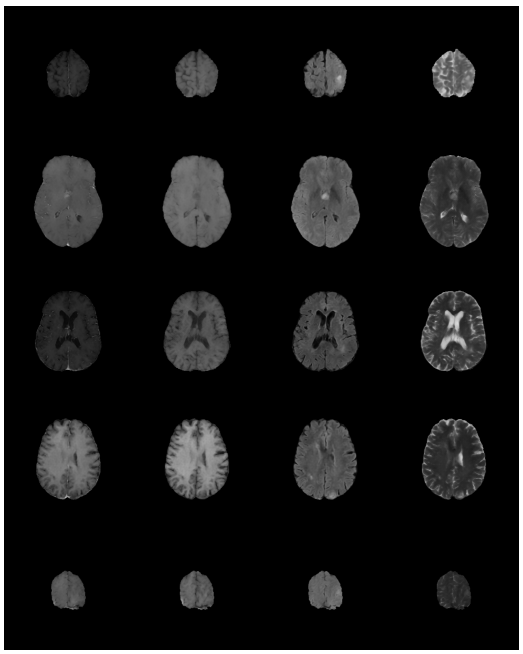
(g) Conditioning: 10 mm



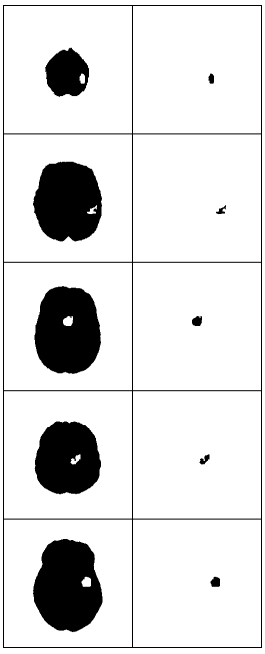
(h) Generated: 10 mm



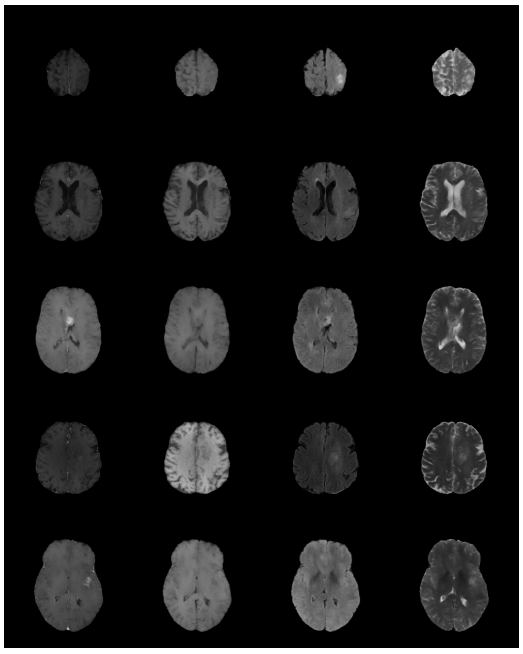
(i) Conditioning: 15 mm



(j) Generated: 15 mm



(k) Conditioning: 20 mm



(l) Generated: 20 mm

Figure B.4: Generated samples from the synthetic dataset using Configuration G4 across different lesion sizes. Subplot columns show the conditioning signal (left) and the corresponding generated sample (right). Within the conditioning signal, the columns correspond to the one-hot encoded brain mask (left) and the optional lesion target (right). Rows correspond to different 2D slices from individual subjects. Within the generated samples, columns correspond to MRI sequences ordered as T_1ce, T_1w, T_2-FLAIR, and T_2w.

Table B.1: Configurations of the generative models shown in Table 4.1. Differences in hyperparameters reflect their consecutive development with an emphasis on reducing complexity.

Hyperparameter	Config. G1	Config. G2	Config. G3	Config. G3wide	Config. G4
Activation Function			SiLU		
Attention Channels per Head			64		
Diffusion Mean Type			$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$		
Diffusion Steps			1000		
Dropout Rate			10		
Learning Rate			0.0001		
Number of Residual Blocks			2		
Output Channels			4		
Patch Size			(256, 256, 1)		
Attention Resolution	(16,)	(32, 16, 8)	(32, 16, 8)	(32, 16, 8)	(32, 16, 8)
Batch Size	10	12	50	50	50
Channel Factor	(1, 1, 2, 3, 4)	(1, 1, 2, 4)	(1, 2, 3, 4)	(1, 2, 4, 8)	(1, 2, 3, 4)
Diffusion Noising	cosine	linear scaled	linear scaled	linear scaled	linear scaled
Diffusion Variance Type	$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$	$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$	$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$	$\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$	β_t
EMA Decay	0.995	0.9999	0.9	0.99	0.9
Hidden Channels	64	320	320	256	256

Table B.2: Hyperparameters of the first-stage models used for evaluating patch-based MRI sequence encoding.

Hyperparameter	**KL-AE**	**VQ-AE**
Batch Size	12	
Channel Factor	(1, 2, 4, 4)	
Discriminator Channels	64	
Discriminator Layers	3	
Discriminator Loss	hinge	
Hidden Channels	128	
Learning Rate	3e-05	
Number of Residual Blocks	2	
Patch Size	(256, 256, 1)	
Pixel Loss	L_1	
Codebook	-	16384×4

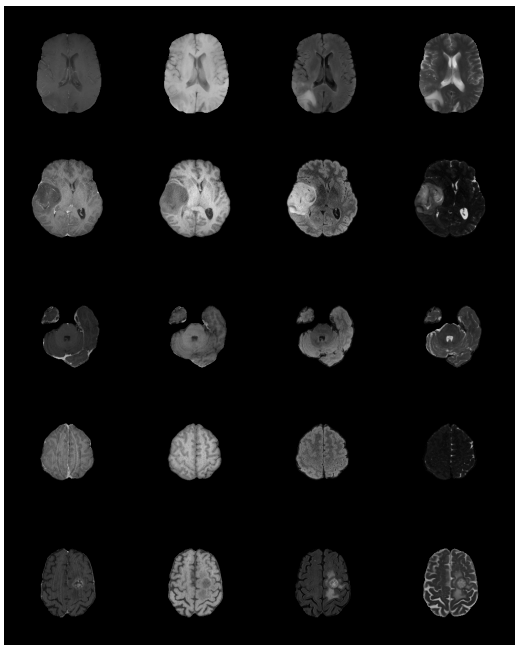
Table B.3: Final hyperparameters of the first-stage model used throughout for patch-based MRI sequence encoding, where "cropped" indicates the patch-based variant.

Hyperparameter	**KL-AE**	**KL-AE (cropped)**
Batch Size	12	
Discriminator Channels	64	
Discriminator Layers	3	
Discriminator Loss	hinge	
Embedding Dimension	4	
Hidden Channels	128	
Learning Rate	3e-05	
Number of Residual Blocks	2	
Patch Size	(256, 256, 1)	
Pixel Loss	L_1	
Channel Factor	(1, 2, 4)	(1, 2, 4, 4)

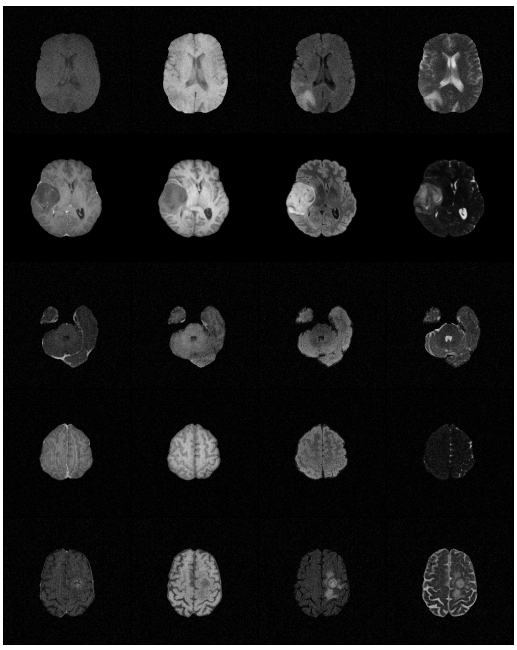
Table B.4: First-stage model configurations for binary lesion mask encoding. "cropped" indicates the patch-based first-stage model.

Hyperparameter	**KL-AE**	**KL-AE (cropped)**
Batch Size	40	
Discriminator Channels	64	
Discriminator Layers	3	
Discriminator Loss	hinge	
Embedding Dimension	4	
Hidden Channels	64	
Learning Rate	0.0001	
Number of Residual Blocks	1	
Patch Size	(256, 256, 1)	
Pixel Loss	L_1	
Channel Factor	(1, 2, 4)	(1, 2, 4, 4)

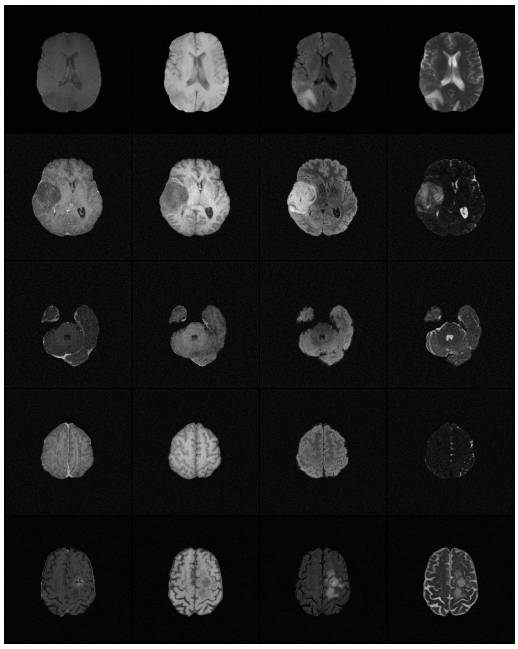
B.2 Detecting small brain tumours



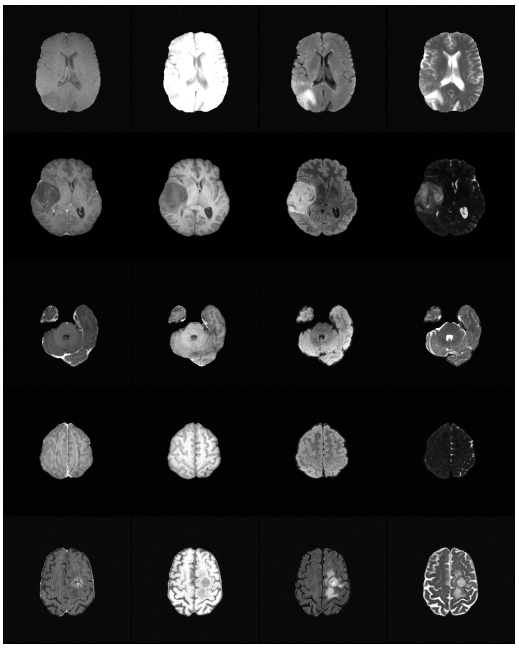
(a) Noise Input



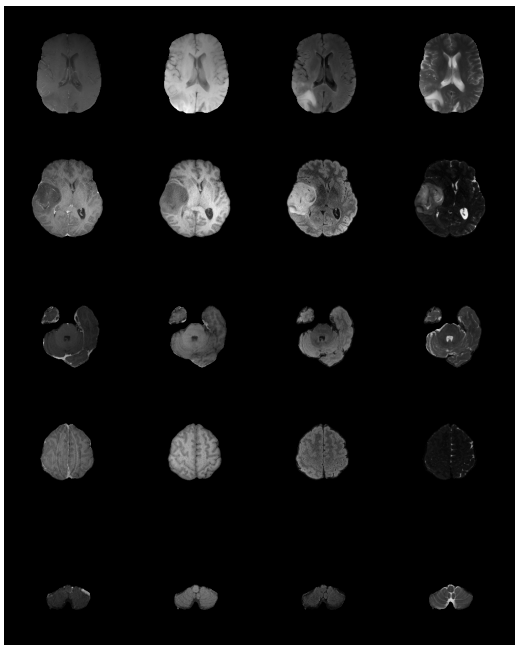
(b) Gaussian Noise



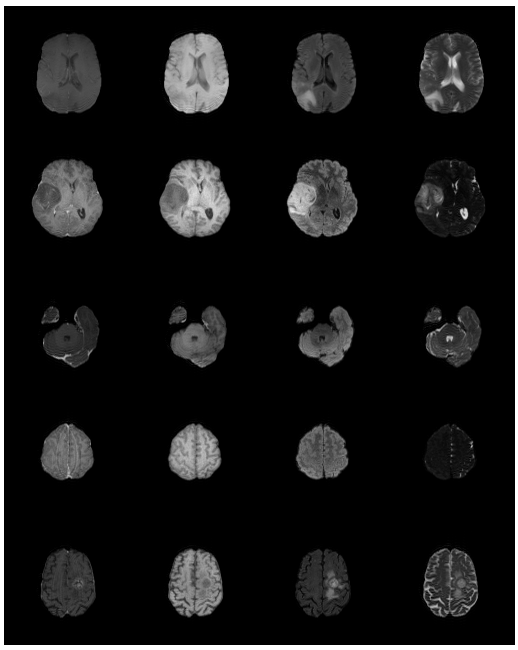
(c) Rician Noise



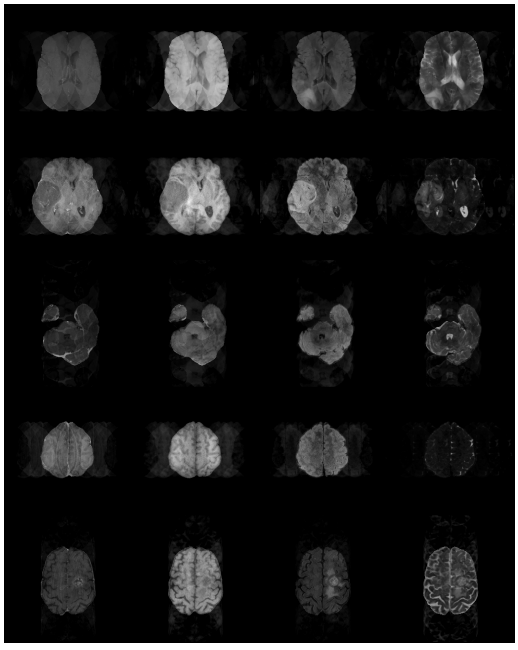
(d) Non-Chi Noise



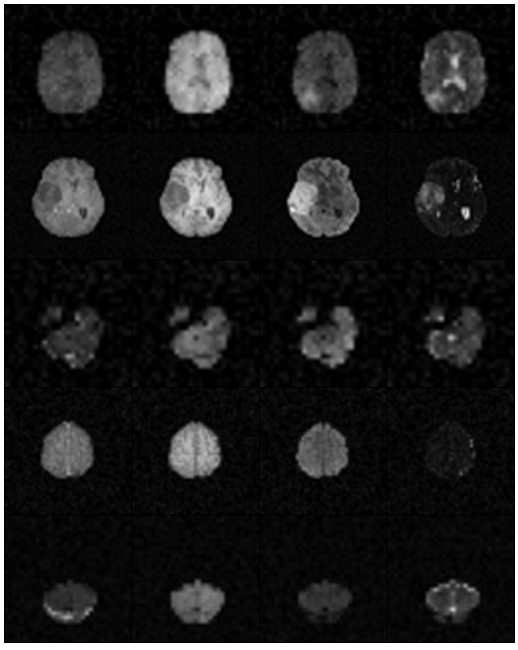
(e) Bias field



(f) Motion artifact



(g) Ghosting



(h) Complete MRI degradation

Figure B.5: Visualisations of MRI degradations, including different noise types, bias fields, motion artefacts, and the full degradation pipeline. Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

Table B.5: Parameters of the degradation pipelines. Parameter ranges are given in brackets. Each pipeline contains $N = 2$ degradation levels, with the parameters for each level specified in parentheses.

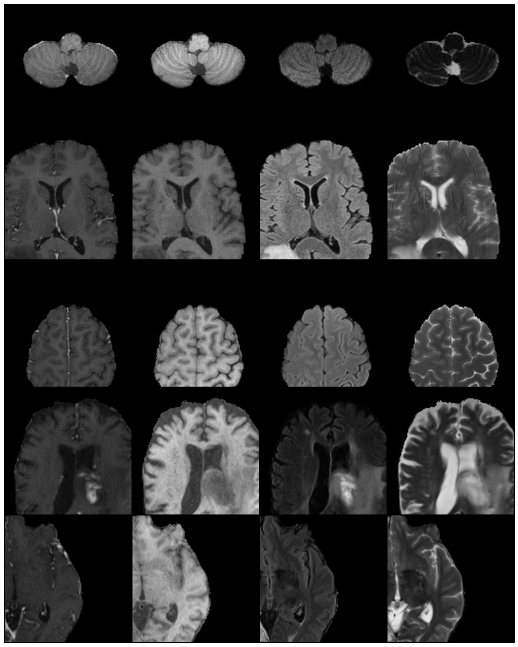
	Parameter	ESRGAN	Adapted for MRI
General	Downscale factor	4	
	No degradation p	0.01	
Reshaping	Resize p	up: (0.2, 0.3), down: (0.7, 0.4), keep: (0.1, 0.3)	
	Resize range	([0.3, 1.5], [0.6, 1.2])	
Blur	p	(1.0, 0.5)	
	σ	([0.2, 1.5], [0.2, 1.0])	
	Iso	0.45	0.40
	Aniso	0.25	0.60
	Generalised Iso	0.12	-
	Generalised Aniso	0.03	-
	Plateau Iso	0.12	-
	Plateau Aniso	0.03	-
Noise	Gaussian p	(0.5, 0.5)	(0.45, 0.45)
	Gaussian σ	([1, 15], [1, 12])	([2, 15], [1, 12])
	Poisson p	(0.5, 0.5)	-
	Poisson σ	([0.05, 2.0], [0.05, 1.0])	-
	Rician p	-	(0.45, 0.45)
	Rician σ	-	([0.01, 0.05], [0.005, 0.02])
	Chi p	-	(0.1, 0.1)
	Chi σ	-	([0.002, 0.02], [0.001, 0.01])
	Chi coil range	-	([1, 3], [1, 3])
B_F	p	-	(0.2, 0.1)
	Range	-	[0.1, 1.0]
Motion	p	-	(0.0, 0.0)
	Rotation range	-	[10, 10]
	Translation range	-	[2, 2]
	Steps	-	[2, 2]

Continued on next page

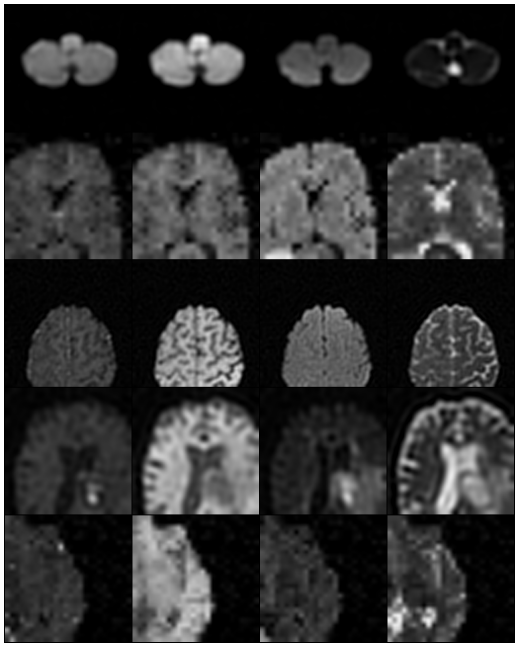
Table B.5: Parameters of the degradation pipelines. Parameter ranges are given in brackets. Each pipeline contains $N = 2$ degradation levels, with the parameters for each level specified in parentheses. (Continued)

Ghosting	p	-	(0.0, 0.0)
	Count range	-	([4, 10], [4, 10])
	Axes	-	([0, 1], [0, 1])
	Intensity range	-	([0.5, 1.0], [0.5, 1.0])

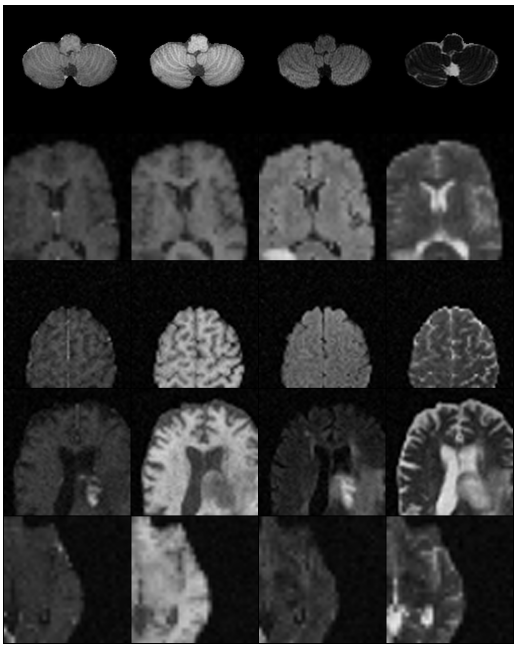
Parameters: p: Probability, σ: Standard deviation
Abbreviations: Iso: Isotropic, Aniso: Anisotropic



(a) Input image



(b) ESRGAN degradation



(c) MRI degradation

Figure B.6: Comparison of degradation pipelines for MRI data. (a) shows the input, (b) the corresponding low-resolution image generated with Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), and (c) the low-resolution image generated with the MRI-specific pipeline. Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w.

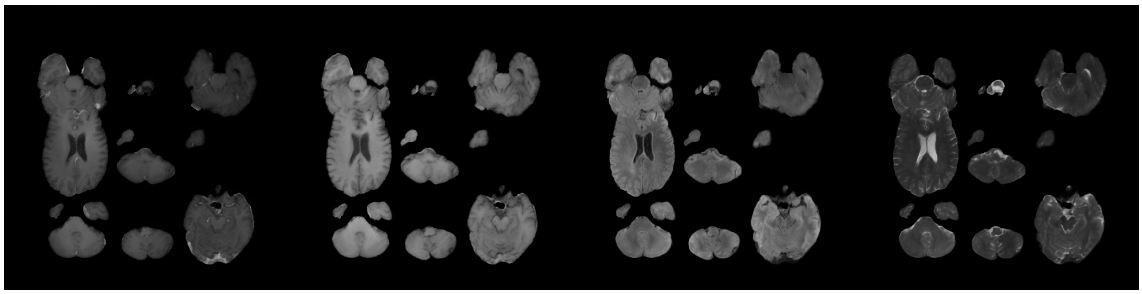


Figure B.7: Patch-based inference using a SR-LDM trained on full images. Each patch produces a brain-like artefact, presumably because the first-stage model maps input patches into a latent space shaped by training on whole-brain anatomy.

Table B.6: Grid search parameters for the anomaly map generation.

Parameter	**Grid Search Space**	**Description**
Number of encoding steps N	400, 500, 600, 700	The encoding and decoding process uses N steps for sampling, rather than traversing the full T steps in each direction.
Classifier-free guidance scale C	3.0, 5.0	Strength of the classifier-free guidance mechanism (see Section 2.3.6 and Equation (2.45))
Timestep aggregation T	[0, 5], [0, 10], [0, 15], [0, 20]	Aggregation of timesteps of the gradient-based signals given by start and stop index.
Aggregation method	Mean, Median, Sum	The method used to aggregate the timesteps of the gradient-based signals and (independently) the channels of the MRI sequences
Morphological methods	Erosion, Dilation, Opening, Closing	Morphological operations applied to the coarse and fine anomaly maps. Up to two operations can be applied in sequence from the set of operations.

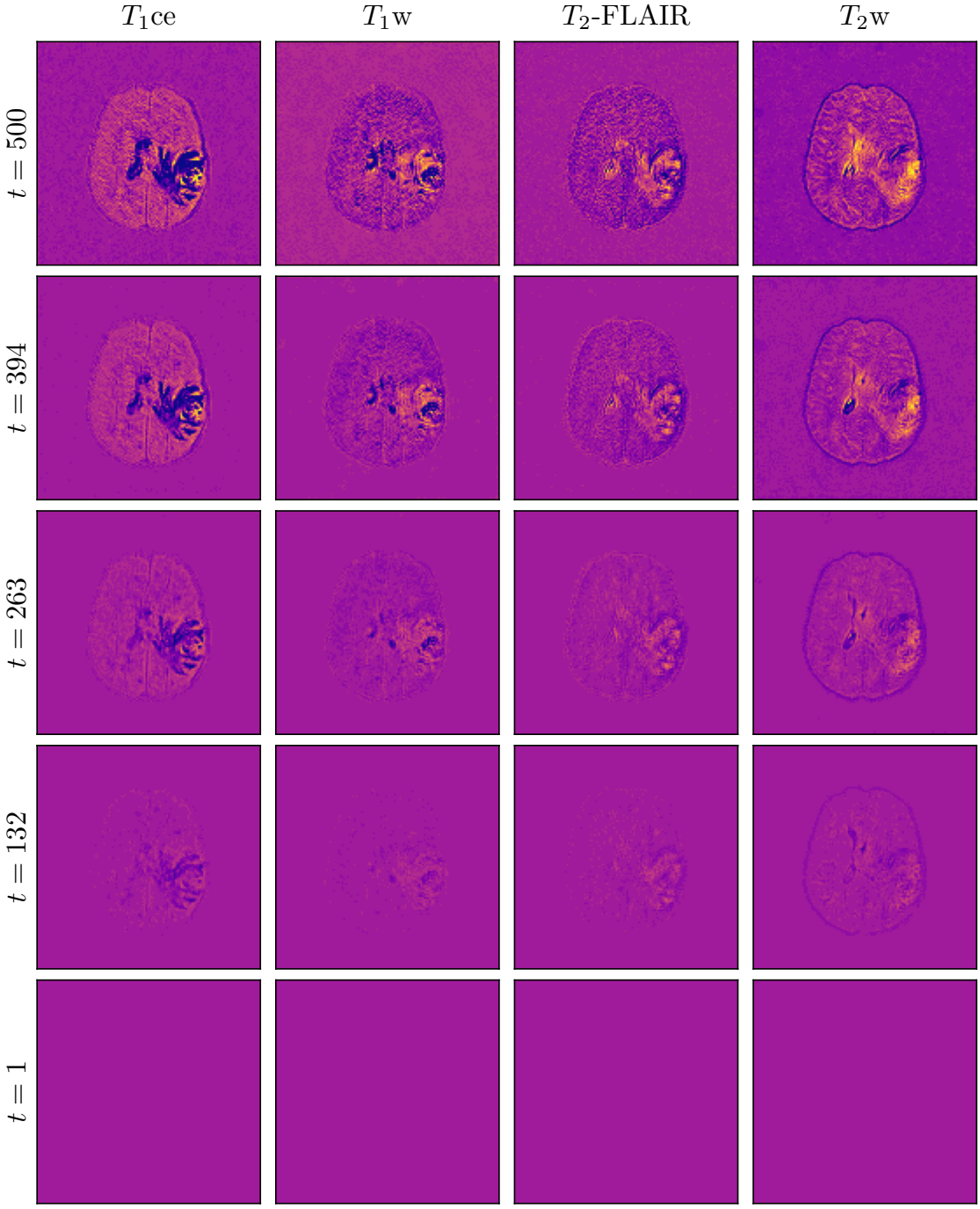


Figure B.8: Visualisation of vanishing gradient across diffusion timesteps, illustrating the effect of temporal distance from the starting point of the reverse process N. The diffusion trajectory begins at $N = 500$ and proceeds backward to $t = 1$, as described in Sections 2.3.1 and 2.3.6. Rows correspond to selected timesteps during sampling, whereas columns indicate the MRI sequence. The chosen colormap enhances visibility of spatial gradient differences.

Table B.7: Configuration of the 2D-LDM for SR evaluation. The patch size denotes the training patch size and does not necessarily correspond to the patch size used during inference.

Hyperparameter	**2D-LDM**
Activation Function	SiLU
Attention Channels per Head	64
Attention Resolution	(32, 16, 8)
Batch Size	24
Channel Factor	(1, 2, 3, 4)
Diffusion Mean Type	$\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$
Diffusion Noising	linear scaled
Diffusion Steps	1000
Diffusion Variance Type	β_t
Dropout Rate	10
EMA Decay	0.9
Hidden Channels	256
Number of Residual Blocks	2
Output Channels	4
Patch Size	(256, 256, 1)

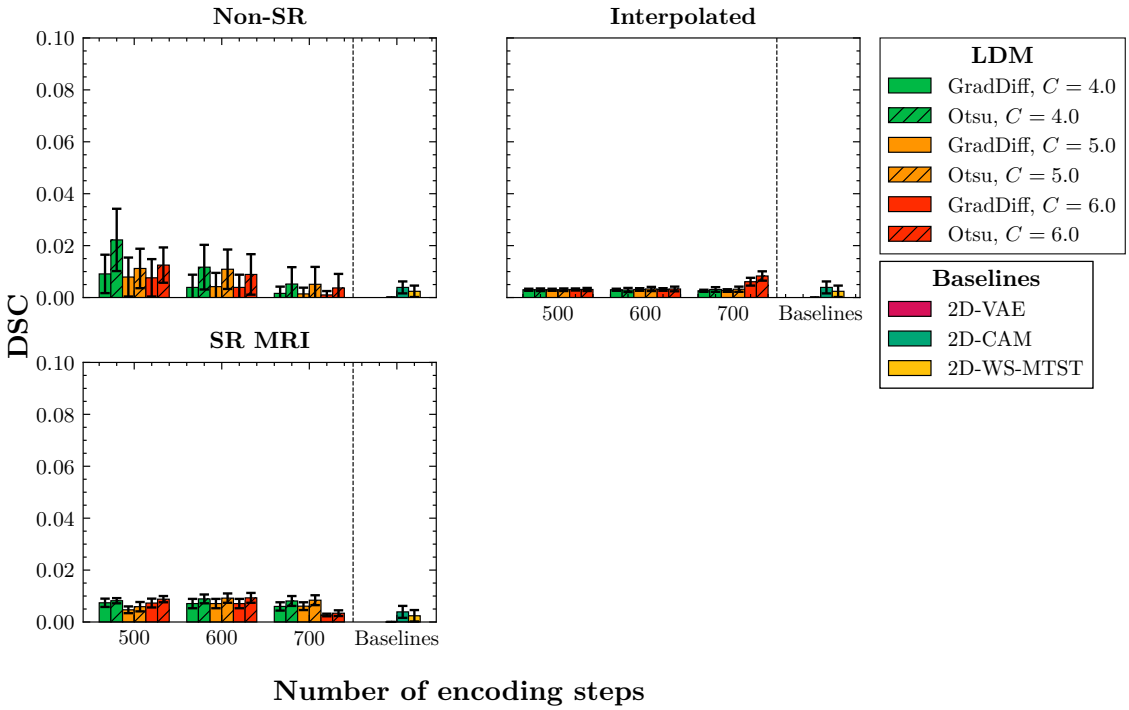


Figure B.9: Performance for the SD dataset with 5 mm lesion diameter. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the Dice similarity coefficient (DSC). The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI degradation pipeline. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

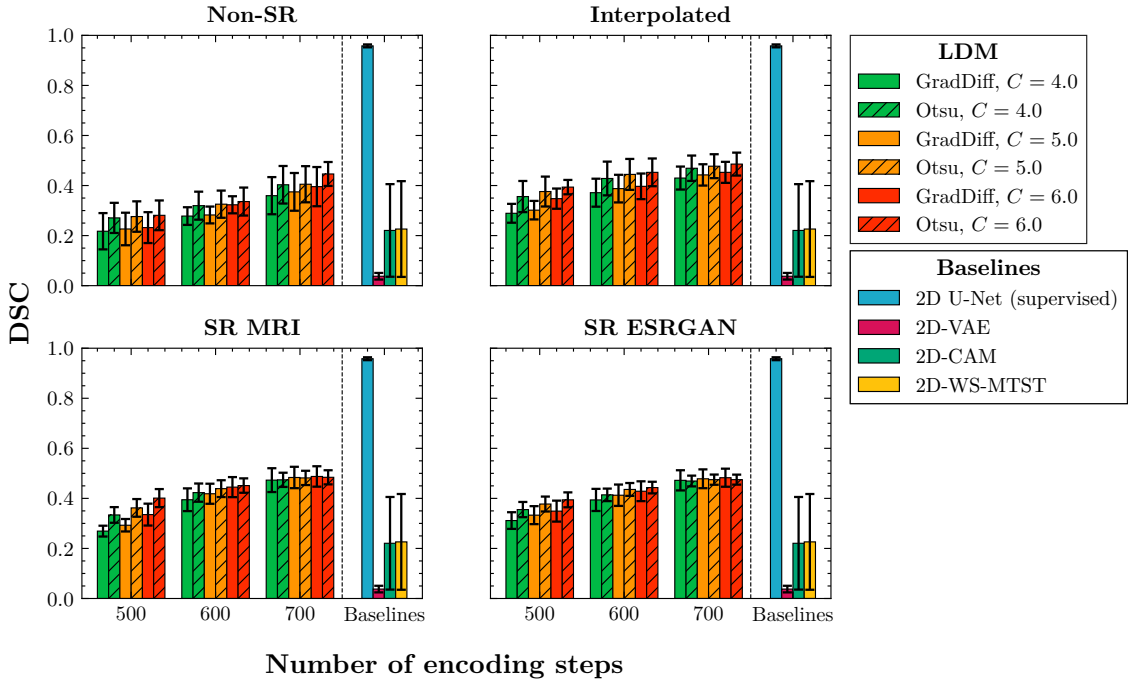


Figure B.10: Performance for the FD dataset including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

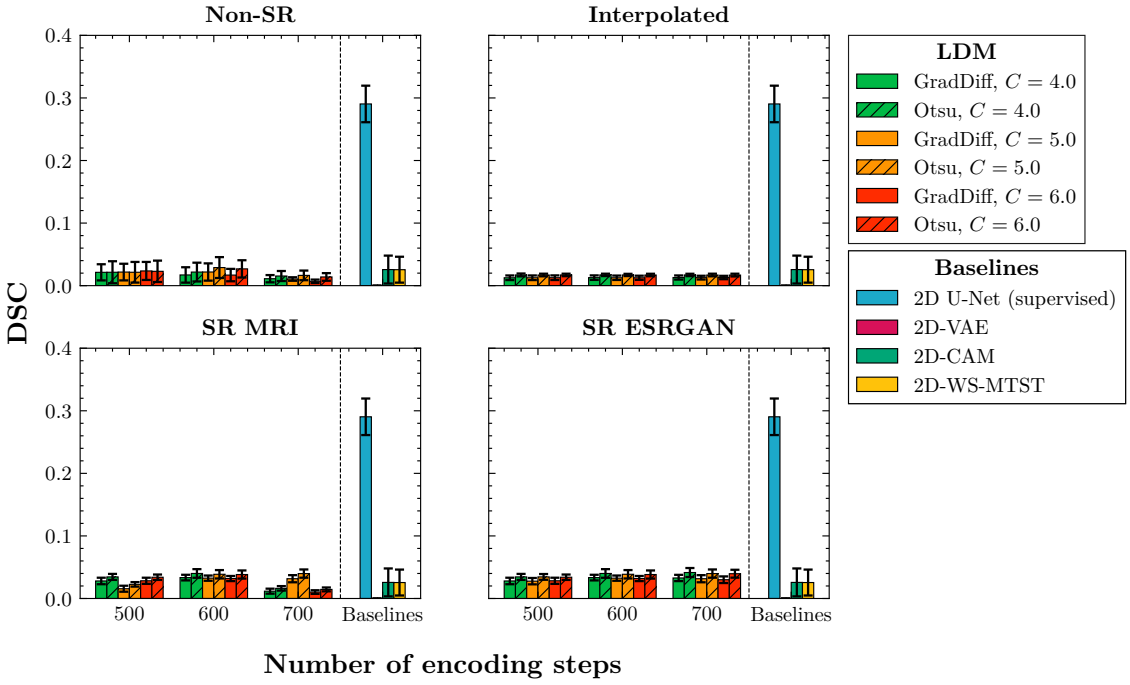


Figure B.11: Performance for the SD dataset with 10 mm lesion diameter, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

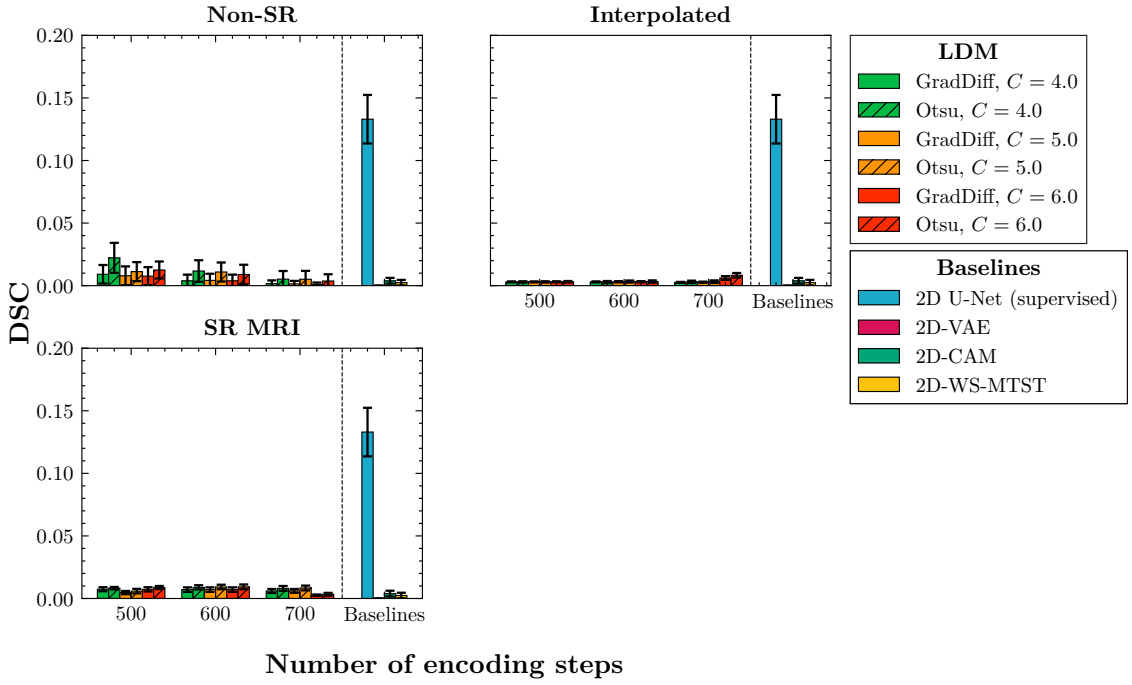


Figure B.12: Performance for the SD dataset with 5 mm lesion diameter, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI degradation pipeline. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

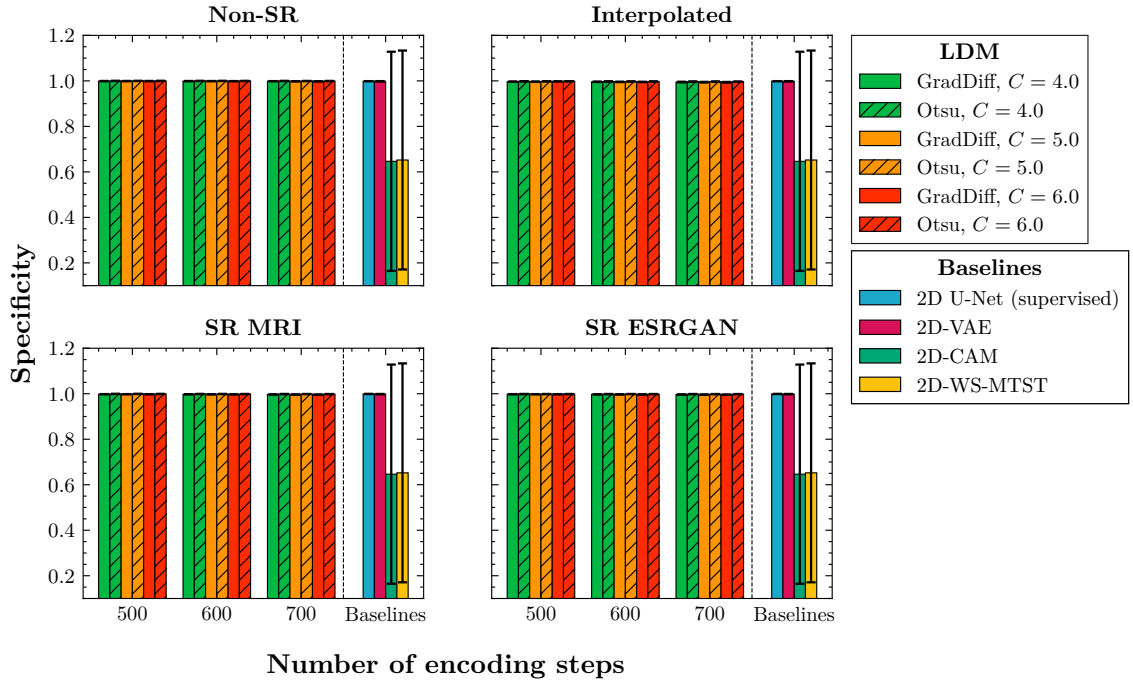


Figure B.13: Specificity for the FD dataset including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the DSC. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

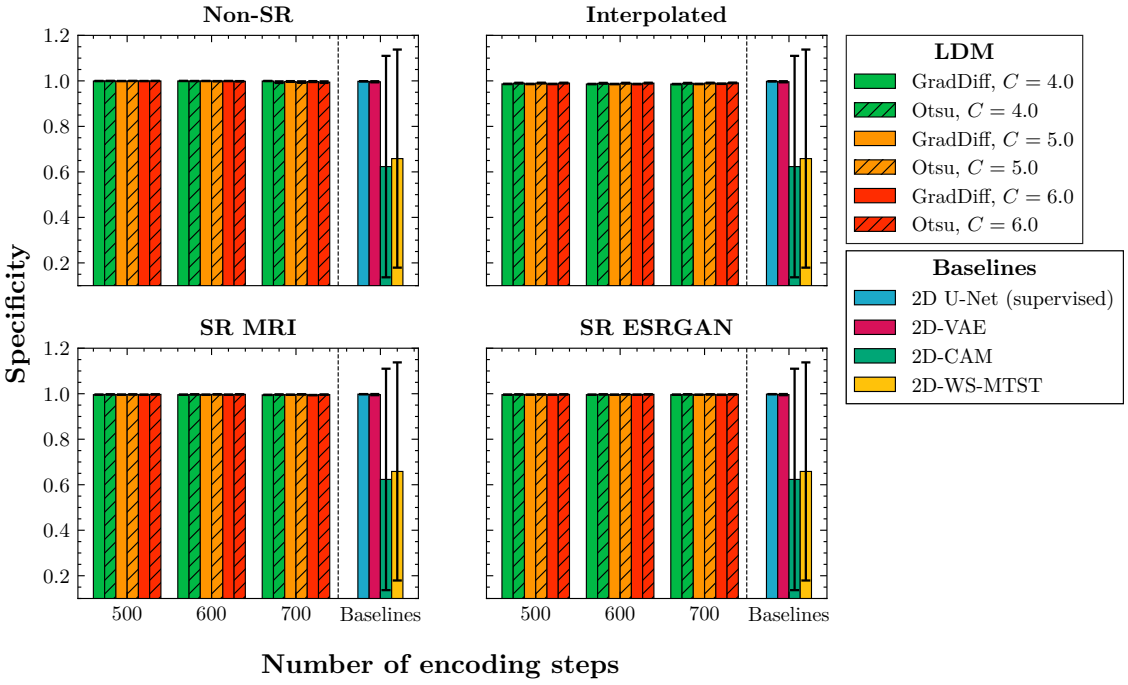


Figure B.14: Specificity for the SD dataset with 10 mm lesion diameter, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the specificity. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI and ESRGAN degradation pipelines. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

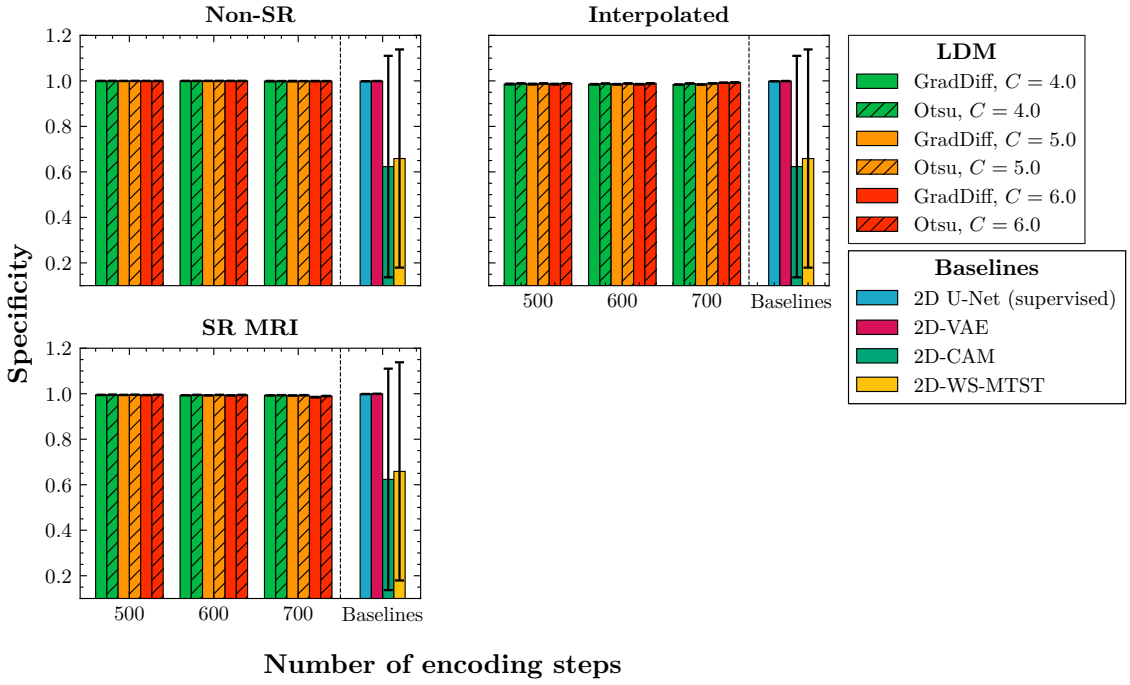
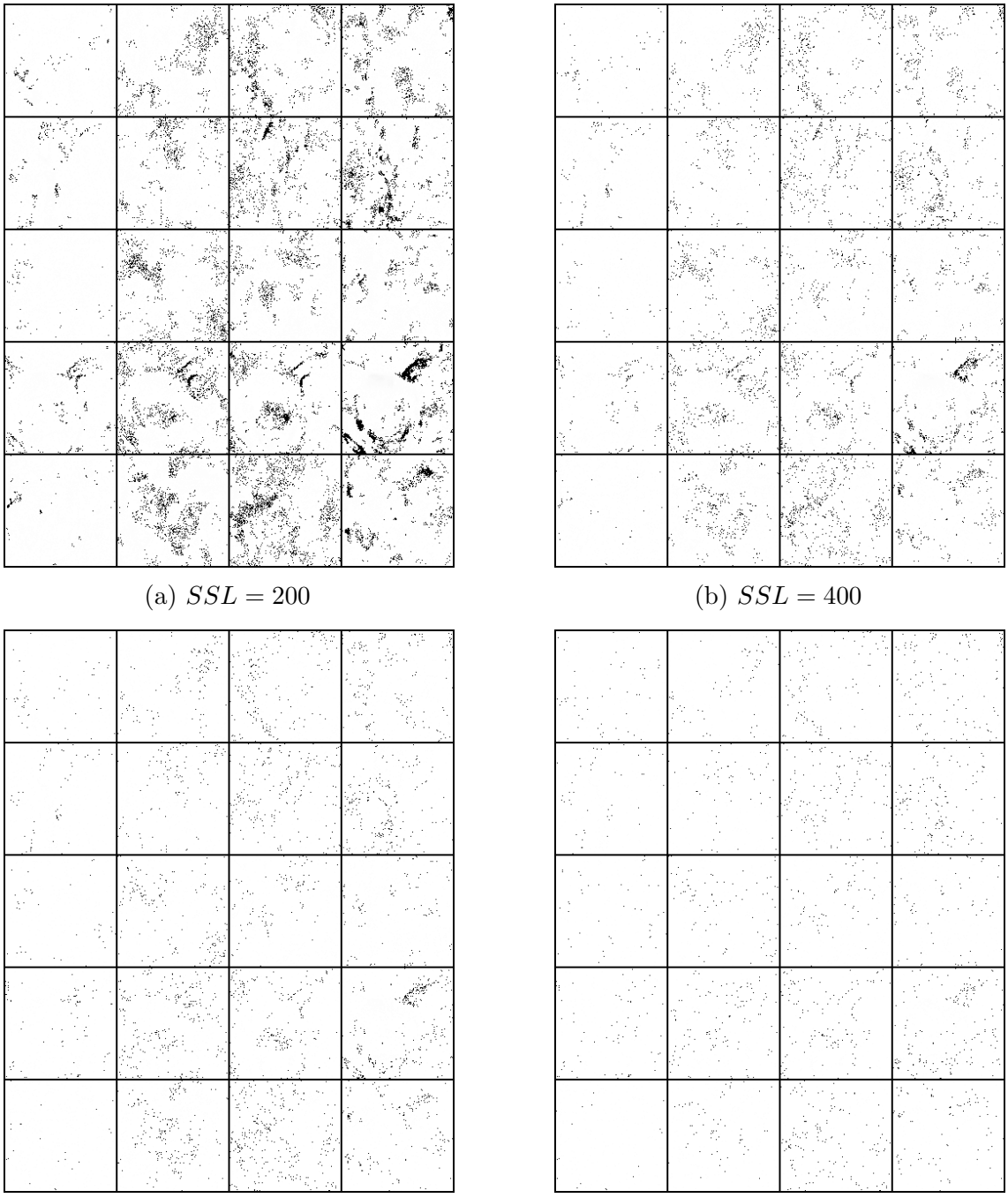


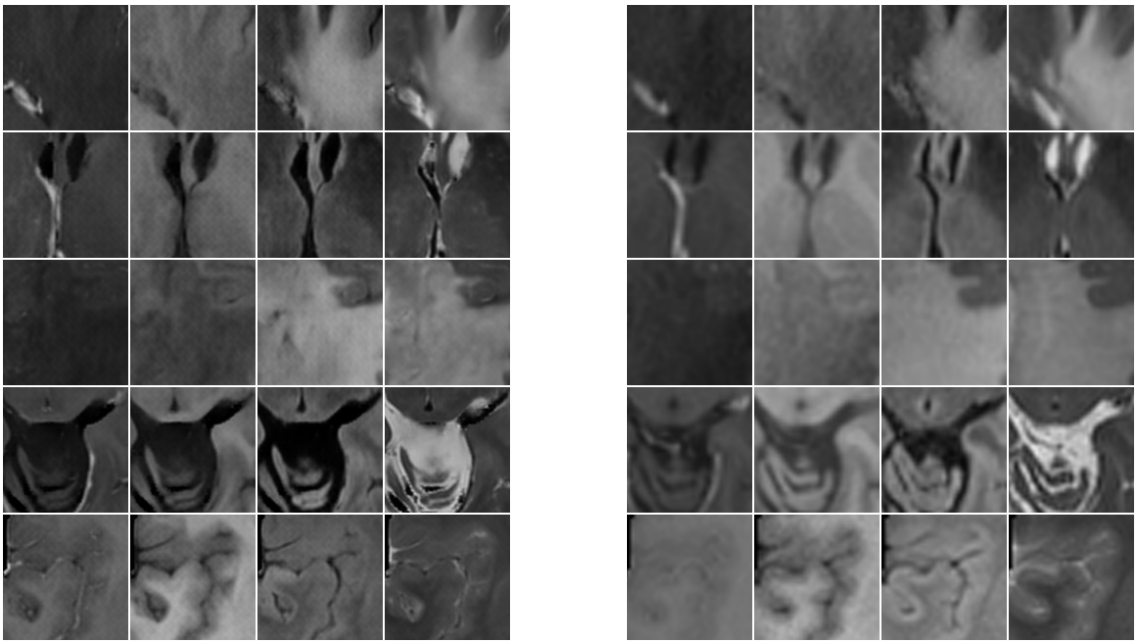
Figure B.15: Specificity for the SD dataset with 5 mm lesion diameter, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis shows the specificity. The top row presents results on non-SR samples, while the bottom row shows 2D-LDM performance on SR data using the MRI degradation pipeline. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.



(a) $SSL = 200$ (b) $SSL = 400$

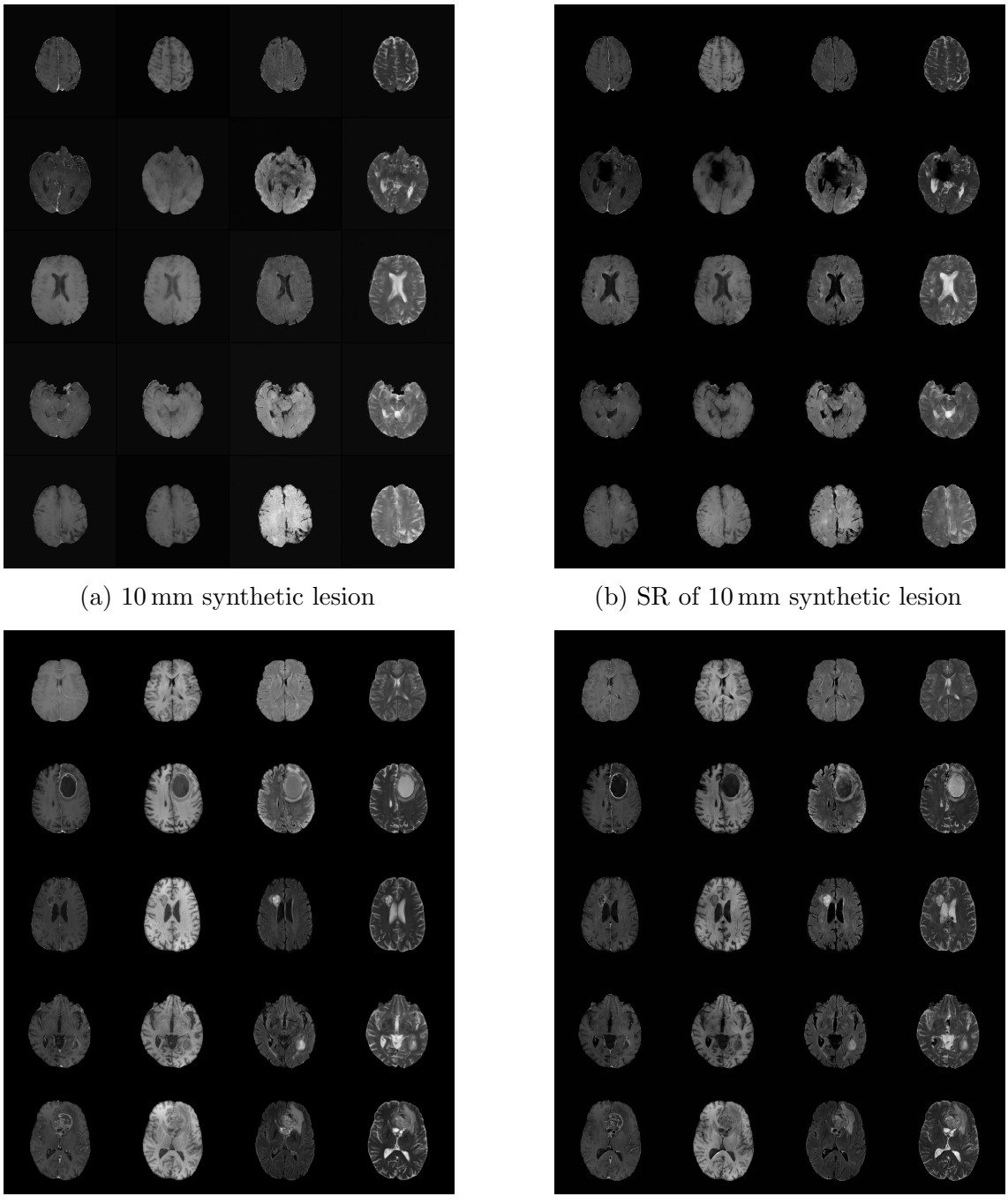
(c) $SSL = 600$ (d) $SSL = 800$

Figure B.16: Absolute difference to the full sequence with $SSL = 1000$ steps for SR. White regions indicate low error, black regions high error. Rows correspond to different 2D slices from individual subjects. The visualisations highlight how SR quality degrades as the diffusion sequence is shortened and more steps are skipped. Results remain largely acceptable with only minor artefacts up to 400-600 steps. *Note:* SSL denotes subsequence length in terms of the shortened sequence introduced by DDIM (J. Song et al., 2021).



(a) SR (b) Upscaled input

Figure B.17: Comparison of SR and conventional interpolation. Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. The results highlight that SR improves fidelity and refines structural details compared to interpolation.



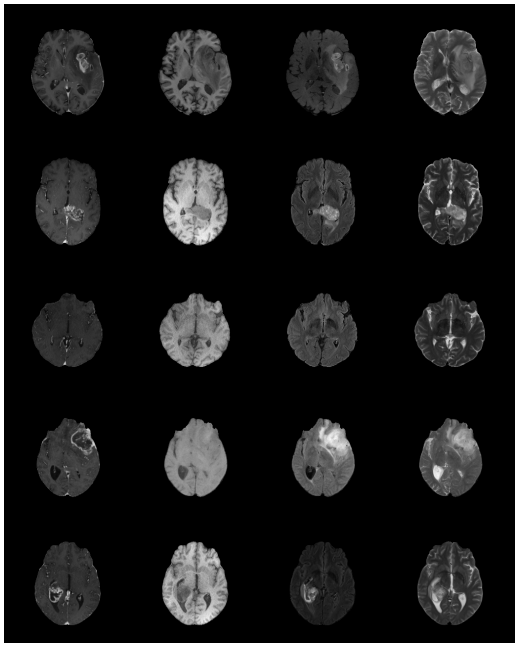
(a) 10 mm synthetic lesion

(b) SR of 10 mm synthetic lesion

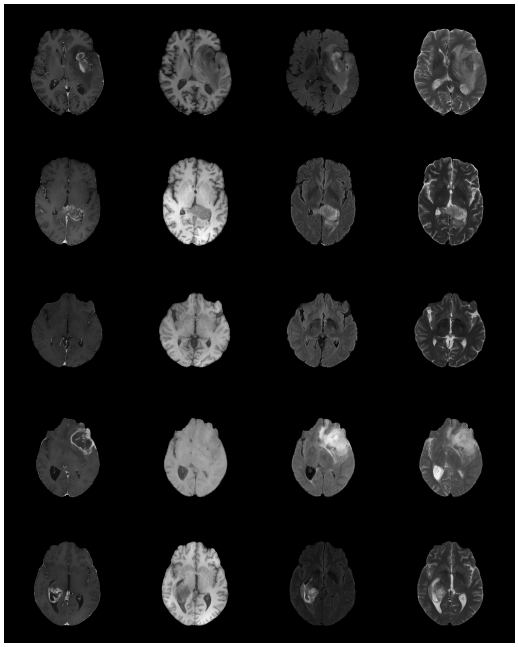
(c) Real lesion

(d) SR of real lesion

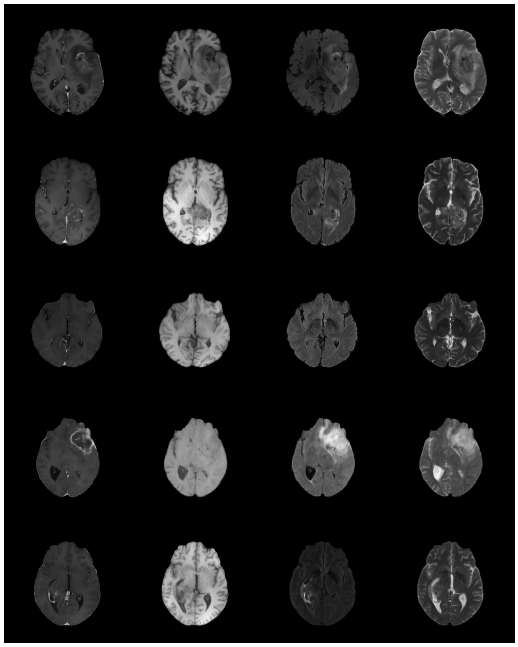
Figure B.18: SR on synthetic small lesions obtained through Configuration G4. The images are generated using the SR model trained on the real adult dataset. Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. Compared to real lesions (bottom row), the synthetic SR outputs (top row) indicate a noticeable distribution shift, reflected in visible artefacts.



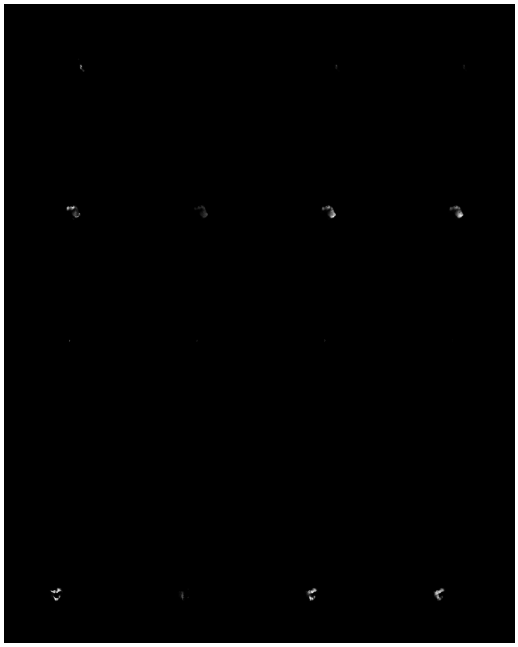
(a) Input



(b) Unconditional SR



(c) Healthy SR counterfactual



(d) Difference between (b) and (c)

Figure B.19: Samples from the unified SR and anomaly detection model. (a) shows the input image, (b) presents the SR result from unconditional sampling, (c) displays the forced healthy counterfactual, and (d) shows the difference between (b) and (c). Rows correspond to different 2D slices from individual subjects. The MRI sequences (columns) are ordered as follows: T_1ce, T_1w, T_2-FLAIR, and T_2w. Conditional sampling alters local image regions but does so inconsistently, underperforming compared to the sequential approach.

Appendix C

Generalisability to paediatric populations

C.1 Pre-trained segmentation model

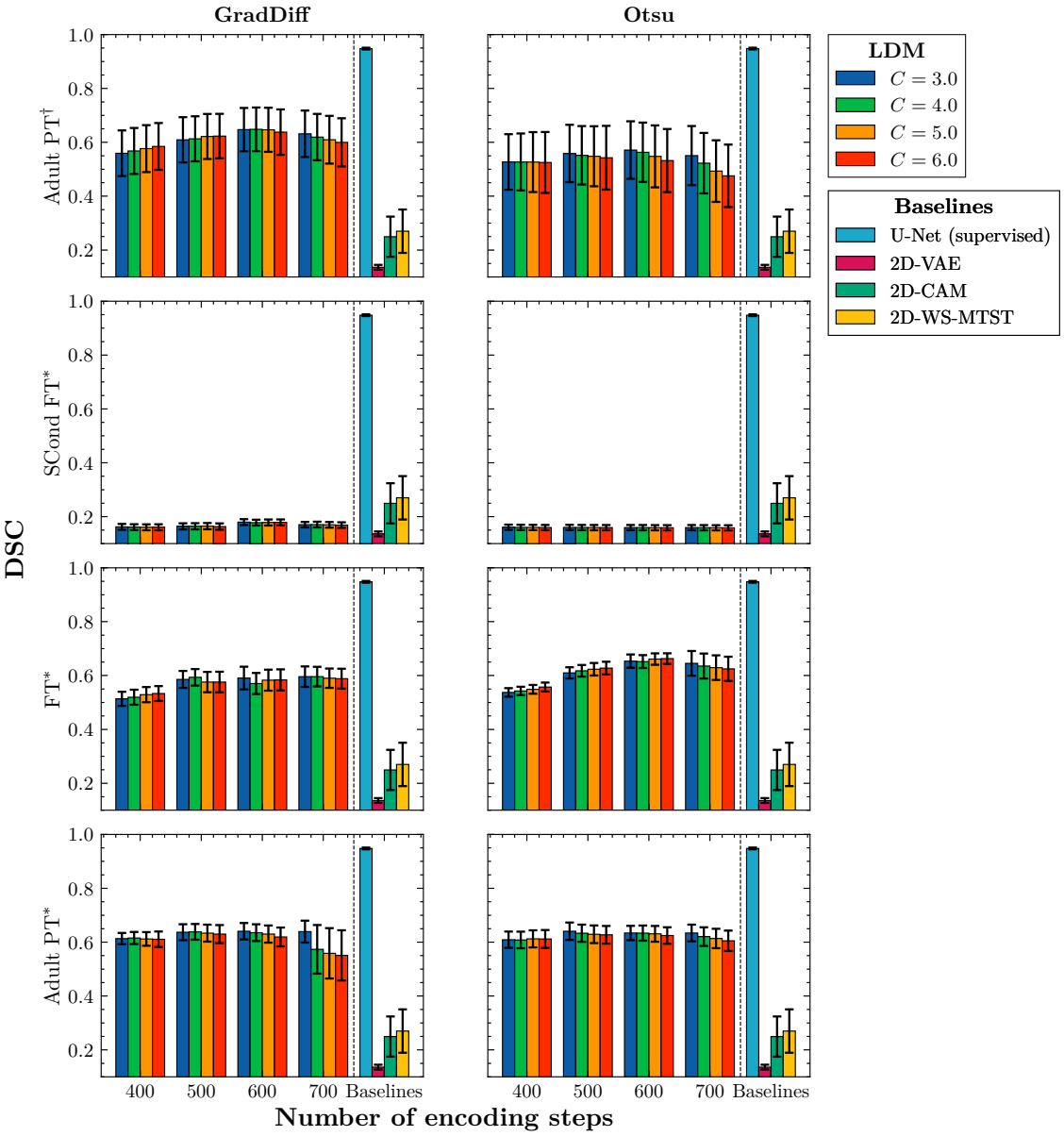


Figure C.1: Paediatric brain tumour segmentation performance measured by DSC including supervised baseline. Each row represents a different training strategy of the 2D-LDM, and each column corresponds to a different thresholding method. The x-axis shows the number of encoding steps N, while the y-axis indicates the DSC score. Colours denote classifier-guidance strength C, and error bars reflect bootstrapped confidence intervals. Baselines are evaluated without fine-tuning and are shown across all plots for visual comparability. PT: pre-trained model; FT: fine-tuned model; SCond: structural encoder conditioning. Superscripts indicate the evaluation set: † refers to the full paediatric test set ($n = 99$), and * refers to the reduced fine-tuning test subset ($n = 20$).

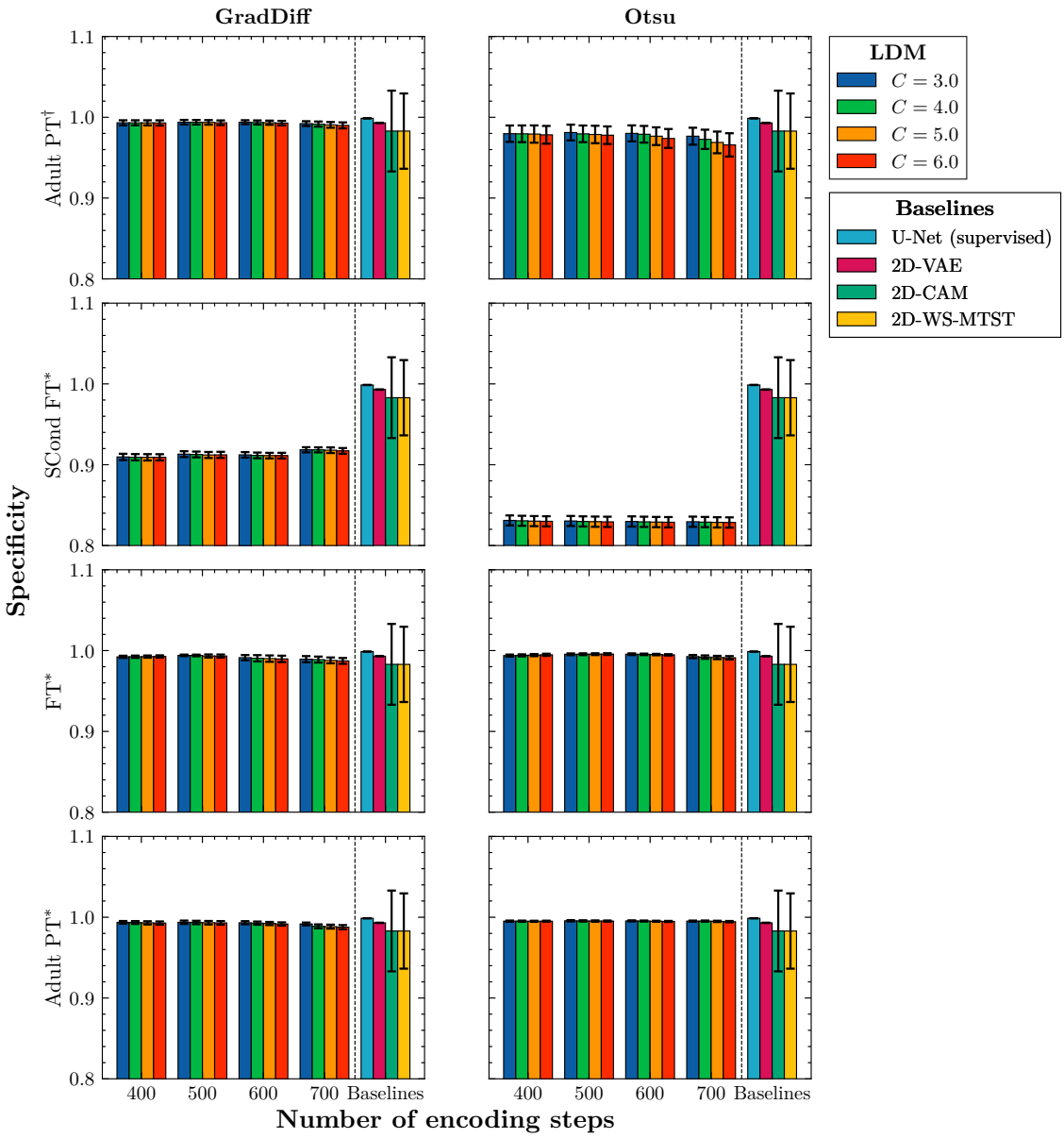


Figure C.2: Paediatric brain tumour segmentation performance measured by specificity including supervised baseline. Each row represents a different training strategy of the 2D-LDM, and each column corresponds to a different thresholding method. The x-axis shows the number of encoding steps N, while the y-axis indicates the specificity. Colours denote classifier-guidance strength C, and error bars reflect bootstrapped confidence intervals. Baselines are evaluated without fine-tuning and are shown across all plots for visual comparability. PT: pre-trained model; FT: fine-tuned model; SCond: structural encoder conditioning. Superscripts indicate the evaluation set: † refers to the full paediatric test set ($n = 99$), and * refers to the reduced fine-tuning test subset ($n = 20$).

C.2 Evaluation on private data collection

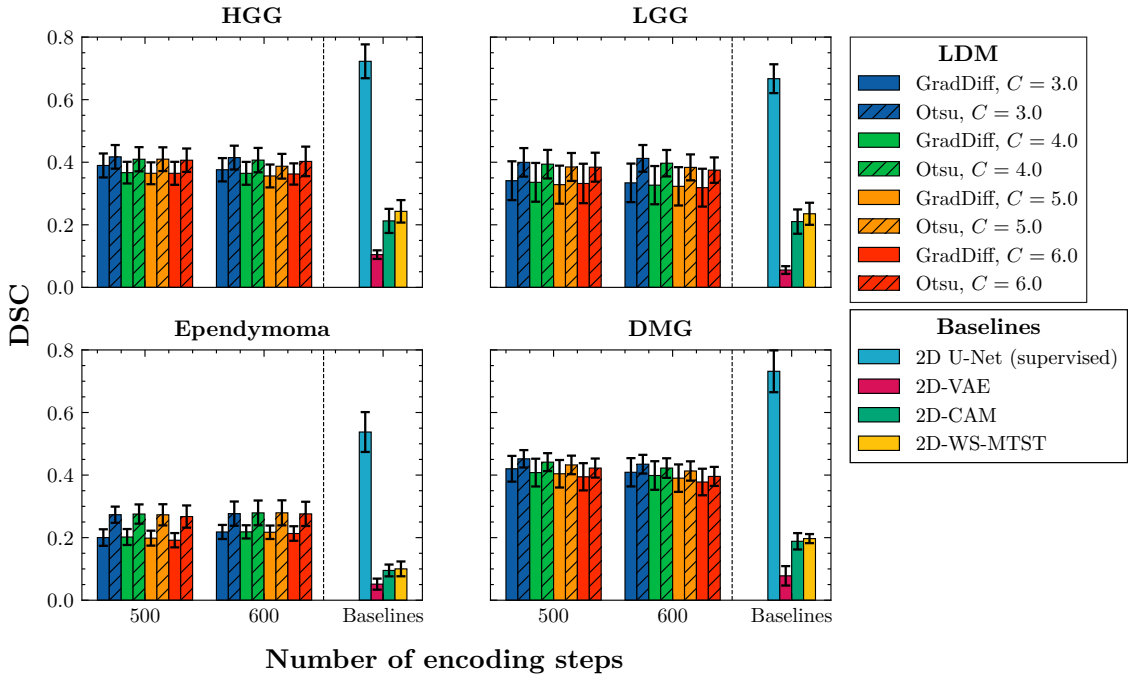


Figure C.3: Quantitative results for the CHOP dataset, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis reports the DSC. Each subplot corresponds to a tumour subtype. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

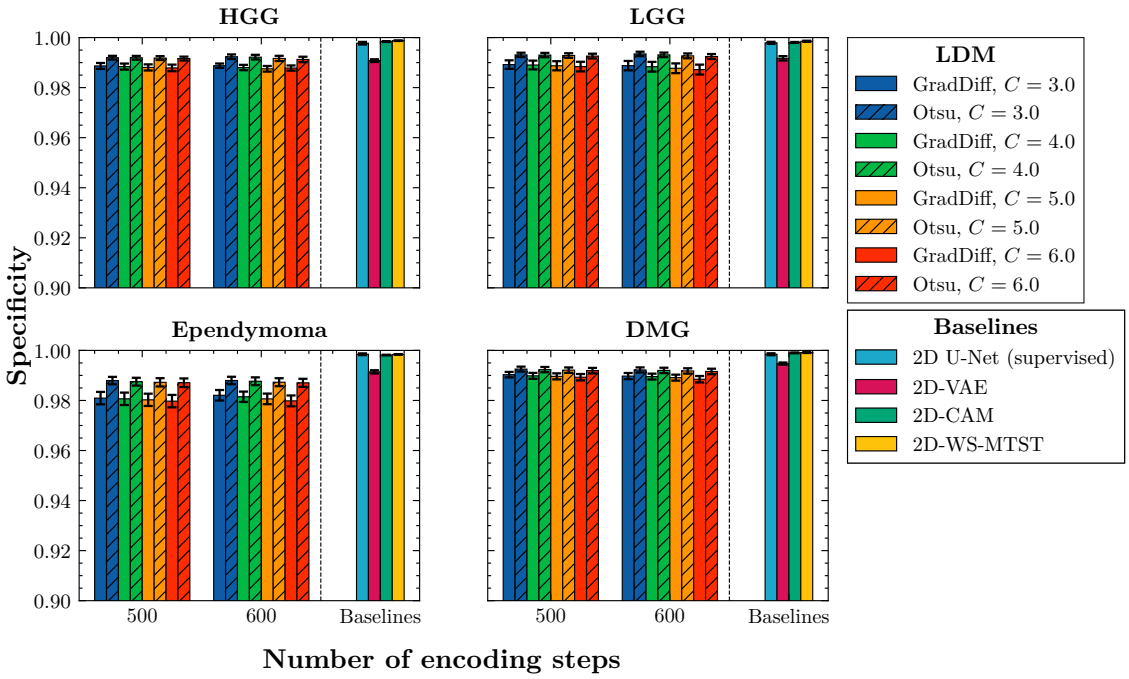


Figure C.4: Specificity for the CHOP dataset, including supervised baseline. The x-axis shows the number of encoding steps N, with baseline results separated by a vertical dotted line. The y-axis reports the specificity. Each subplot corresponds to a tumour subtype. Bar colours represent the diffusion gradient scale C and distinguish baselines from diffusion-based models, while textures denote the applied thresholding method. Error bars indicate bootstrapped confidence intervals.

Bibliography

Abraham, N., & Khan, N. M. (2019). A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. ***2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)***, 683–687. https://doi.org/10.1109/ISBI.2019.8759329

Adewole, M., Rudie, J. D., Gbdamosi, A., Toyobo, O., Raymond, C., Zhang, D., Omidiji, O., Akinola, R., Suwaid, M. A., Emegoakor, A., Ojo, N., Aguh, K., Kalaiwo, C., Babatunde, G., Ogunleye, A., Gbadamosi, Y., Iorpagher, K., Calabrese, E., Aboian, M., ... Anazodo, U. C. (2023). The Brain Tumor Segmentation (BraTS) Challenge 2023: Glioma Segmentation in Sub-Saharan Africa Patient Population (BraTS-Africa). ***ArXiv***, arXiv:2305.19369v1.

Agravat, R. R., & Raval, M. S. (2020). Brain Tumor Segmentation and Survival Prediction. ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries***, 338–348. https://doi.org/10.1007/978-3-030-46640-4_32

Aja-Fernández, S., & Vegas-Sánchez-Ferrero, G. (2016). ***Statistical Analysis of Noise in MRI***. Springer International Publishing. https://doi.org/10.1007/978-3-319-39934-8

AlAmir, M., & AlGhamdi, M. (2022). The Role of Generative Adversarial Network in Medical Image Analysis: An in-depth survey. ***ACM Computing Surveys***, 3527849. https://doi.org/10.1145/3527849

Alemany, M., Velasco, R., Simó, M., & Bruna, J. (2021). Late effects of cancer treatment: Consequences for long-term brain cancer survivors. ***Neuro-Oncology Practice***, ***8***(1), 18–30. https://doi.org/10.1093/nop/npaa039

Ali, S., Ghatwary, N., Jha, D., Isik-Polat, E., Polat, G., Yang, C., Li, W., Galdran, A., Ballester, M.-Á. G., Thambawita, V., Hicks, S., Poudel, S., Lee, S.-W., Jin, Z., Gan, T., Yu, C., Yan, J., Yeo, D., Lee, H., ... East, J. E. (2024). Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. ***Scientific Reports***, ***14***(1), 2032. https://doi.org/10.1038/s41598-024-52063-x

Amian, M., & Soltaninejad, M. (2020). Multi-resolution 3D CNN for MRI Brain Tumor Segmentation and Survival Prediction. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 221–230, Vol. 11992). Springer International Publishing. https://doi.org/10.1007/978-3-030-46640-4_21

An, J., Wendt, L., Wiese, G., Herold, T., Rzepka, N., Mueller, S., Koch, S. P., Hoffmann, C. J., Harms, C., & Boehm-Sturm, P. (2023). Deep learning-based automated lesion segmentation on mouse stroke magnetic resonance images. ***Scientific Reports***, ***13***(1), 13341. https://doi.org/10.1038/s41598-023-39826-8

Ardalan, Z., & Subbian, V. (2022). Transfer Learning Approaches for Neuroimaging Analysis: A Scoping Review. ***Frontiers in Artificial Intelligence***, ***5***, 780405. https://doi.org/10.3389/frai.2022.780405

Armstrong, G. T. (2010). Long-term survivors of childhood central nervous system malignancies: The experience of the Childhood Cancer Survivor Study. ***European journal of paediatric neurology: EJPN: official journal of the European Paediatric Neurology Society***, ***14***(4), 298–303. https://doi.org/10.1016/j.ejpn.2009.12.006

Atlason, H. E., M.d, A. L., Sigurdsson, S., M.d, V. G., & Ellingsen, L. M. (2019). Unsupervised brain lesion segmentation from MRI using a convolutional autoencoder. ***Medical Imaging 2019: Image Processing***, ***10949***, 372–378. https://doi.org/10.1117/12.2512953

Avesta, A., Hossain, S., Lin, M., Aboian, M., Krumholz, H. M., & Aneja, S. (2023). Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-

Segmentation. ***Bioengineering***, ***10***(2), 181. https://doi.org/10.3390/bioengineering10020181

Axel, L., Summers, R. M., Kressel, H. Y., & Charles, C. (1986). Respiratory effects in two-dimensional Fourier transform MR imaging. ***Radiology***, ***160***(3), 795–801. https://doi.org/10.1148/radiology.160.3.3737920

Ayana, G., Dese, K., Abagaro, A. M., Jeong, K. C., Yoon, S.-D., & Choe, S.-w. (2024). Multistage transfer learning for medical images. ***Artificial Intelligence Review***, ***57***(9), 1–47. https://doi.org/10.1007/s10462-024-10855-7

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., … Bakas, S. (2021). ***The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification***.

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), 170117. https://doi.org/10.1038/sdata.2017.117

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., … Menze, B. (with Apollo-University Of Cambridge Repository & University Of Cambridge). (2018). Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. https://doi.org/10.17863/CAM.38755

Banerjee, A., & Nicolaides, T. (2017). Low-Grade Gliomas. In ***Pediatric CNS Tumors*** (pp. 1–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-30789-3

Bangalore Yogananda, C. G., Wagner, B., Nalawade, S. S., Murugesan, G. K., Pinho, M. C., Fei, B., Madhuranthakam, A. J., & Maldjian, J. A. (2020). Fully Automated Brain Tumor Segmentation and Survival Prediction of Gliomas Using Deep Learning and MRI. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 99–112, Vol. 11993). Springer International Publishing. https://doi.org/10.1007/978-3-030-46643-5_10

Barratt, S., & Sharma, R. (2018). ***A Note on the Inception Score***. arXiv: 1801.01973 [stat]. https://doi.org/10.48550/arXiv.1801.01973

Bauer, S., May, C., Dionysiou, D., Stamatakos, G., Büchler, P., & Reyes, M. (2012). Multiscale modeling for image analysis of brain tumor studies. ***IEEE transactions on bio-medical engineering***, ***59***(1), 25–29. https://doi.org/10.1109/TBME.2011.2163406

Baur, C., Denner, S., Wiestler, B., Navab, N., & Albarqouni, S. (2021). Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. ***Medical Image Analysis***, ***69***, 101952. https://doi.org/10.1016/j.media.2020.101952

Baur, C., Wiestler, B., Albarqouni, S., & Navab, N. (2019). Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. https://doi.org/10.1007/978-3-030-11723-8_16

Behjati, S., Gilbertson, R. J., & Pfister, S. M. (2021). Maturation Block in Childhood Cancer. ***Cancer Discovery***, ***11***(3), 542–544. https://doi.org/10.1158/2159-8290.CD-20-0926

Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., & Schlaefer, A. (2024). Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI. ***Medical Imaging with Deep Learning***, 1019–1032.

Bengtsson, M., Keles, E., Durak, G., Anwar, S., Velichko, Y. S., Linguraru, M. G., Waanders, A. J., & Bagci, U. (2025). A New Logic for Pediatric Brain Tumor Segmentation. ***2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)***, 1–5. https://doi.org/10.1109/ISBI60581.2025.10980809

Bercea, C. I., Wiestler, B., Rueckert, D., & Schnabel, J. A. (2024). Diffusion Models with Implicit Guidance for Medical Anomaly Detection. ***Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024***, ***LNCS 15011***, 211–220.

Bernhardt, M., Castro, D. C., Tanno, R., Schwaighofer, A., Tezcan, K. C., Monteiro, M., Bannur, S., Lungren, M. P., Nori, A., Glocker, B., Alvarez-Valle, J., & Oktay, O. (2022). Active label cleaning for improved dataset quality under resource constraints. ***Nature Communications***, ***13***(1), 1161. https://doi.org/10.1038/s41467-022-28818-3

Bhattacharya, M., Gupta, S., Singh, A., Chen, C., Singh, G., & Prasanna, P. (2025). ***BrainMRDiff: A Diffusion Model for Anatomically Consistent Brain MRI Synthesis***. arXiv: 2504.04532 [eess]. https://doi.org/10.48550/arXiv.2504.04532

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. ***CA: A Cancer Journal for Clinicians***, ***69***(2), 127–157. https://doi.org/10.3322/caac.21552

Bishop, C. M., & Bishop, H. (2024a). Convolutional Networks. In ***Deep Learning*** (pp. 287–324). Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4_10

Bishop, C. M., & Bishop, H. (2024b). The Deep Learning Revolution. In ***Deep Learning*** (pp. 1–22). Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4_1

Bishop, C. M., & Bishop, H. (2024c). ***Deep learning: Foundations and concepts***. Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4

Bishop, C. M., & Bishop, H. (2024d). Diffusion Models. In ***Deep Learning*** (pp. 581–607). Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4_20

Bishop, C. M., & Bishop, H. (2024e). Transformers. In ***Deep Learning*** (pp. 357–406). Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4_12

Borji, A. (2022). Pros and cons of GAN evaluation measures: New developments. ***Computer Vision and Image Understanding***, ***215***, 103329. https://doi.org/10.1016/j.cviu.2021.103329

Boutry, N., Chazalon, J., Puybareau, E., Tochon, G., Talbot, H., & Géraud, T. (2020). Using Separated Inputs for Multimodal Brain Tumor Segmentation with 3D U-Net-like Architectures. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 187–199, Vol. 11992). Springer International Publishing. https://doi.org/10.1007/978-3-030-46640-4_18

Bria, A., Marrocco, C., & Tortorella, F. (2020). Addressing class imbalance in deep learning for small lesion detection on medical images. ***Computers in Biology and Medicine***, ***120***, 103735. https://doi.org/10.1016/j.compbiomed.2020.103735

Bruno, M. A., Walker, E. A., & Abujudeh, H. H. (2015). Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. ***RadioGraphics***, ***35***(6), 1668–1676. https://doi.org/10.1148/rg.2015150023

Burkett, B. J., Fagan, A. J., Felmlee, J. P., Black, D. F., Lane, J. I., Port, J. D., Rydberg, C. H., & Welker, K. M. (2021). Clinical 7-T MRI for neuroradiology: Strengths, weaknesses, and ongoing challenges. ***Neuroradiology***, ***63***(2), 167–177. https://doi.org/10.1007/s00234-020-02629-z

Buzuti, L. F., & Thomaz, C. E. (2023). Fréchet AutoEncoder Distance: A new approach for evaluation of Generative Adversarial Networks. ***Computer Vision and Image Understanding***, ***235***, 103768. https://doi.org/10.1016/j.cviu.2023.103768

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., ... Feng, A.

(2022). ***MONAI: An open-source framework for deep learning in healthcare***. arXiv: 2211.02701 [cs]. https://doi.org/10.48550/arXiv.2211.02701

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). Extracting training data from large language models. ***30th USENIX Security Symposium (USENIX Security 21)***, 2633–2650.

Carrete, L. R., Young, J. S., & Cha, S. (2022). Advanced Imaging Techniques for Newly Diagnosed and Recurrent Gliomas. ***Frontiers in Neuroscience***, ***16***, 787755. https://doi.org/10.3389/fnins.2022.787755

Central nervous system tumours (5th ed). (2021). International agency for research on cancer.

Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2023). Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. ***Medical Image Analysis***, ***87***, 102792. https://doi.org/10.1016/j.media.2023.102792

Chang, E. L., Hassenbusch, S. J., Shiu, A. S., Lang, F. F., Allen, P. K., Sawaya, R., & Maor, M. H. (2003). The role of tumor size in the radiosurgical management of patients with ambiguous brain metastases. ***Neurosurgery***, ***53***(2), 272–280, discussion 280–281. https://doi.org/10.1227/01.neu.0000073546.61154.9a

Chang, J. S., Haas-Kogan, D. A., & Mueller, S. (2017). High-Grade Gliomas. In ***Pediatric CNS Tumors*** (pp. 37–50). Springer International Publishing. https://doi.org/10.1007/978-3-319-30789-3

Chen, C.-F. R., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. ***2021 IEEE/CVF International Conference on Computer Vision (ICCV)***, 347–356. https://doi.org/10.1109/ICCV48922.2021.00041

Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., & Gao, W. (2021). Pre-Trained Image Processing Transformer, 12294–12305. https://doi.org/10.1109/CVPR46437.2021.01212

Chen, H., An, J., Jiang, B., Xia, L., Bai, Y., & Gao, Z. (2023). WS-MTST: Weakly Supervised Multi-Label Brain Tumor Segmentation With Transformers. ***IEEE journal of biomedical and health informatics***, ***27***(12), 5914–5925. https://doi.org/10.1109/JBHI.2023.3321602

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Self-supervised learning for medical image analysis using image context restoration. ***Medical Image Analysis***, ***58***, 101539. https://doi.org/10.1016/j.media.2019.101539

Chen, M., Wang, P., Guo, Y., Yin, Y., Wang, L., Su, Y., & Gong, G. (2022). The effect of time delay for magnetic resonance contrast-enhanced scan on imaging for small-volume brain metastases. ***NeuroImage : Clinical***, ***36***, 103223. https://doi.org/10.1016/j.nicl.2022.103223

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. (2022). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions.

Chen, W., Zhou, W., Zhu, L., Cao, Y., Gu, H., & Yu, B. (2022). MTDCNet: A 3D multi-threading dilated convolutional network for brain tumor automatic segmentation. ***Journal of Biomedical Informatics***, ***133***, 104173. https://doi.org/10.1016/j.jbi.2022.104173

Chen, X., Williams, B. M., Vallabhaneni, S. R., Czanner, G., Williams, R., & Zheng, Y. (2019). Learning Active Contour Models for Medical Image Segmentation. ***2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 11624–11632. https://doi.org/10.1109/CVPR.2019.01190

Chen, Z., Tian, Z., Zhu, J., Li, C., & Du, S. (2022). C-CAM: Causal CAM for Weakly Supervised Semantic Segmentation on Medical Image, 11676–11685.

Cheplygina, V. (2019). Cats or CAT scans: Transfer learning from natural or medical image source data sets? ***Current Opinion in Biomedical Engineering***, ***9***, 21–27. https://doi.org/10.1016/j.cobme.2018.12.005

Cheplygina, V., de Bruijne, M., & Pluim, J. P. (2019). Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. ***Medical Image Analysis***, ***54***, 280–296. https://doi.org/10.1016/j.media.2019.03.009

Chong, M. J., & Forsyth, D. (2020). Effectively Unbiased FID and Inception Score and Where to Find Them. ***2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 6069–6078. https://doi.org/10.1109/CVPR42600.2020.00611

Choong, J., & Hameed, N. (2021). Extending Upon a Transfer Learning Approach for Brain Tumour Segmentation. In ***Applied Intelligence and Informatics*** (pp. 60–69, Vol. 1435). Springer International Publishing. https://doi.org/10.1007/978-3-030-82269-9_5

Chung, H., Lee, E. S., & Ye, J. C. (2023). MR Image Denoising and Super-Resolution Using Regularized Reverse Diffusion. ***IEEE Transactions on Medical Imaging***, ***42***(4), 922–934. https://doi.org/10.1109/TMI.2022.3220681

Chung, H., Sim, B., & Ye, J. C. (2022). Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction. ***2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 12403–12412. https://doi.org/10.1109/CVPR52688.2022.01209

Chung, W. J., Chung, H. W., Shin, M. J., Lee, S. H., Lee, M. H., Lee, J. S., Kim, M.-J., & Lee, W. K. (2012). MRI to differentiate benign from malignant soft-tissue tumours of the extremities: A simplified systematic imaging approach using depth, size and heterogeneity of signal intensity. ***The British Journal of Radiology***, ***85***(1018), e831–e836. https://doi.org/10.1259/bjr/27487871

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In ***Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016*** (pp. 424–432, Vol. 9901). Springer International Publishing. https://doi.org/10.1007/978-3-319-46723-8_49

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. ***Journal of Digital Imaging***, ***26***(6), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding, 3213–3223.

Crespi, L., Loiacono, D., & Sartori, P. (2022). Are 3D better than 2D Convolutional Neural Networks for Medical Imaging Semantic Segmentation? ***2022 International Joint Conference on Neural Networks (IJCNN)***, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892850

d'Amati, A., Bargiacchi, L., Rossi, S., Carai, A., Bertero, L., Barresi, V., Errico, M. E., Buccoliero, A. M., Asioli, S., Marucci, G., Del Baldo, G., Mastronuzzi, A., Miele, E., D'Antonio, F., Schiavello, E., Biassoni, V., Massimino, M., Gessi, M., Antonelli, M., & Gianno, F. (2024). Pediatric CNS tumors and 2021 WHO classification: What do oncologists need from pathologists? ***Frontiers in Molecular Neuroscience***, ***17***. https://doi.org/10.3389/fnmol.2024.1268038

Dai, T., Cai, J., Zhang, Y., Xia, S.-T., & Zhang, L. (2019). Second-Order Attention Network for Single Image Super-Resolution. ***2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 11057–11066. https://doi.org/10.1109/CVPR.2019.01132

Dale, B. M., Brown, M. A., & Semelka, R. C. (2015). ***MRI Basic Principles and Applications*** (1st ed.). Wiley. https://doi.org/10.1002/9781119013068

Dar, S. U. H., Seyfarth, M., Ayx, I., Papavassiliu, T., Schoenberg, S. O., Siepmann, R. M., Laqua, F. C., Kahmann, J., Frey, N., Baeßler, B., Foersch, S., Truhn, D., Kather, J. N., & Engelhardt, S. (2025). Unconditional latent diffusion models memorize patient imaging data. ***Nature Biomedical Engineering***. https://doi.org/10.1038/s41551-025-01468-8

Davatzikos, C., Rathore, S., Bakas, S., Pati, S., Bergman, M., Kalarot, R., Sridharan, P., Gastounioti, A., Jahani, N., Cohen, E., Akbari, H., Tunc, B., Doshi, J., Parker, D., & Hsieh, M. (2018). Cancer imaging phenomics toolkit: Quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. ***Journal of Medical Imaging***, ***5***(01), 1. https://doi.org/10.1117/1.JMI.5.1.011018

de Ruiter, M. A., van Mourik, R., Schouten-van Meeteren, A. Y. N., Grootenhuis, M. A., & Oosterlaan, J. (2013). Neurocognitive consequences of a paediatric brain tumour and its treatment: A meta-analysis. ***Developmental Medicine and Child Neurology***, ***55***(5), 408–417. https://doi.org/10.1111/dmcn.12020

Deike-Hofmann, K., Thünemann, D., Breckwoldt, M. O., Schwarz, D., Radbruch, A., Enk, A., Bendszus, M., Hassel, J., Schlemmer, H.-P., & Bäumer, P. (2018). Sensitivity of different MRI sequences in the early detection of melanoma brain metastases. ***PLOS ONE***, ***13***(3), e0193946. https://doi.org/10.1371/journal.pone.0193946

Dempsey, M. F., Condon, B. R., & Hadley, D. M. (2005). Measurement of Tumor "Size" in Recurrent Malignant Glioma: 1D, 2D, or 3D? ***AJNR: American Journal of Neuroradiology***, ***26***(4), 770–776.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. ***2009 IEEE Conference on Computer Vision and Pattern Recognition***, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Detlefsen, N., Borovec, J., Schock, J., Jha, A., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., & Falcon, W. (2022). TorchMetrics - Measuring Reproducibility in PyTorch. ***Journal of Open Source Software***, ***7***(70), 4101. https://doi.org/10.21105/joss.04101

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. ***Proceedings of the 35th International Conference on Neural Information Processing Systems***, 8780–8794.

Diaz, O., Kushibar, K., Osuala, R., Linardos, A., Garrucho, L., Igual, L., Radeva, P., Prior, F., Gkontra, P., & Lekadir, K. (2021). Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. ***Physica Medica***, ***83***, 25–37. https://doi.org/10.1016/j.ejmp.2021.02.007

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: Non-linear independent components estimation. ***3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings***.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). Density estimation using Real NVP. ***Conference Track Proceedings***.

Dobrovoljac, M., Hengartner, H., Boltshauser, E., & Grotzer, M. A. (2002). Delay in the diagnosis of paediatric brain tumours. ***European Journal of Pediatrics***, ***161***(12), 663–667. https://doi.org/10.1007/s00431-002-1088-4

Dockhorn, T., Cao, T., Vahdat, A., & Kreis, K. (2023). Differentially private diffusion models. ***Transactions on Machine Learning Research***.

Dong, H., Yu, F., Jiang, H., Zhang, H., Dong, B., Li, Q., & Zhang, L. (2020). Annotation-Free Gliomas Segmentation Based on a Few Labeled General Brain Tumor Images. ***2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)***, 354–358. https://doi.org/10.1109/ISBI45749.2020.9098366

Dorjsembe, Z., Pao, H.-K., Odonchimed, S., & Xiao, F. (2024). Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. ***IEEE Journal of Biomedical and Health Informatics***, 1–10. https://doi.org/10.1109/JBHI.2024.3385504

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. ***International Conference on Learning Representations***.

Downie, J., Schmidt, M., Kenny, N., D'Arcy, R., Hadskis, M., & Marshall, J. (2007). Paediatric MRI Research Ethics: The Priority Issues. ***Journal of Bioethical Inquiry***, ***4***(2), 85–91. https://doi.org/10.1007/s11673-007-9046-5

Dowson, D. C., & Landau, B. V. (1982). The Fréchet distance between multivariate normal distributions. ***Journal of Multivariate Analysis***, ***12***(3), 450–455. https://doi.org/10.1016/0047-259X(82)90077-X

Drai, M., Testud, B., Brun, G., Hak, J.-F., Scavarda, D., Girard, N., & Stellmann, J.-P. (2022). Borrowing strength from adults: Transferability of AI algorithms for paediatric brain and tumour segmentation. ***European Journal of Radiology***, ***151***, 110291. https://doi.org/10.1016/j.ejrad.2022.110291

Eisenhauer, E. A., Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D., & Verweij, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). ***European Journal of Cancer (Oxford, England: 1990)***, ***45***(2), 228–247. https://doi.org/10.1016/j.ejca.2008.10.026

El Emam, K. (2013). ***Guide to the de-identification of personal health information***. CRC Press/Taylor & Francis Group.

Elazab, A., Abdulazeem, Y. M., Anter, A. M., Hu, Q., Wang, T., & Lei, B. (2018). Macroscopic Cerebral Tumor Growth Modeling From Medical Images: A Review. ***IEEE Access***, ***6***, 30663–30679. https://doi.org/10.1109/ACCESS.2018.2839681

Elazab, A., Wang, C., Gardezi, S. J. S., Bai, H., Hu, Q., Wang, T., Chang, C., & Lei, B. (2020). GP-GAN: Brain tumor growth prediction using stacked 3D generative adversarial networks from longitudinal MR Images. ***Neural Networks***, ***132***, 321–332. https://doi.org/10.1016/j.neunet.2020.09.004

Erdur, A. C., Scholz, D., Buchner, J. A., Combs, S. E., Rueckert, D., & Peeken, J. C. (2024). All Sizes Matter: Improving Volumetric Brain Segmentation on Small Lesions. ***Brain Tumor Segmentation, and Cross-Modality***

Domain Adaptation for Medical Image Segmentation, 177–189. https://doi.org/10.1007/978-3-031-76163-8_16

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)***, 12873–12883.

Falcon, W., & The PyTorch Lightning team. (2019). ***PyTorch lightning*** (Version 1.4). https://doi.org/10.5281/zenodo.3828935

Fang, K., & Li, W.-J. (2020). DMNet: Difference Minimization Network for Semi-supervised Segmentation in Medical Images. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2020***, 532–541. https://doi.org/10.1007/978-3-030-59710-8_52

Feng, X., Tustison, N. J., Patel, S. H., & Meyer, C. H. (2020). Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. ***Frontiers in Computational Neuroscience***, ***14***, 25. https://doi.org/10.3389/fncom.2020.00025

Feng, Y., Cao, Y., An, D., Liu, P., Liao, X., & Yu, B. (2024). DAUnet: A U-shaped network combining deep supervision and attention for brain tumor segmentation. ***Knowledge-Based Systems***, ***285***, 111348. https://doi.org/10.1016/j.knosys.2023.111348

Fernandez, V., Pinaya, W. H. L., Borges, P., Graham, M. S., Tudosiu, P.-D., Vercauteren, T., & Cardoso, M. J. (2024). Generating multi-pathological and multi-modal images and labels for brain MRI. ***Medical Image Analysis***, ***97***, 103278. https://doi.org/10.1016/j.media.2024.103278

Fernandez, V., Pinaya, W. H. L., Borges, P., Tudosiu, P.-D., Graham, M. S., Vercauteren, T., & Cardoso, M. J. (2022). Can Segmentation Models Be Trained with Fully Synthetically Generated Data? ***Simulation and Synthesis in Medical Imaging***, 79–90. https://doi.org/10.1007/978-3-031-16980-9_8

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., & Fookes, C. (2021). Deep Learning for Medical Anomaly Detection - A Survey. ***ACM Computing Surveys***, ***54***(7), 141:1–141:37. https://doi.org/10.1145/3464423

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. ***Engineering Applications of Artificial Intelligence***, ***92***, 103678. https://doi.org/10.1016/j.engappai.2020.103678

Fontanella, A., Mair, G., Wardlaw, J., Trucco, E., & Storkey, A. (2024). Diffusion Models for Counterfactual Generation and Anomaly Detection in Brain Images. ***IEEE Transactions on Medical Imaging***, 1–1. https://doi.org/10.1109/TMI.2024.3460391

Forst, D. A., Nahed, B. V., Loeffler, J. S., & Batchelor, T. T. (2014). Low-Grade Gliomas. ***The Oncologist***, ***19***(4), 403–413. https://doi.org/10.1634/theoncologist.2013-0345

Fritsche, M., Gu, S., & Timofte, R. (2019). Frequency Separation for Real-World Super-Resolution. ***2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)***, 3599–3608. https://doi.org/10.1109/ICCVW.2019.00445

Fry, C. W., Perrow, R., & Paul, S. P. (2014). Brain tumours in children: Importance of early identification. ***British Journal of Nursing***, ***23***(22), 1202–1207. https://doi.org/10.12968/bjon.2014.23.22.1202

Ganatra, H. A. (2025). Machine Learning in Pediatric Healthcare: Current Trends, Challenges, and Future Directions. ***Journal of Clinical Medicine***, ***14***(3), 807. https://doi.org/10.3390/jcm14030807

Ghaffari, M., Samarasinghe, G., Jameson, M., Aly, F., Holloway, L., Chlap, P., Koh, E.-S., Sowmya, A., & Oliver, R. (2022). Automated post-operative brain tumour segmentation: A deep learning model based on transfer learning from pre-operative images. ***Magnetic Resonance Imaging***, ***86***, 28–36. https://doi.org/10.1016/j.mri.2021.10.012

Goldman, R. D., Cheng, S., & Cochrane, D. D. (2017). Improving diagnosis of pediatric central nervous system tumours: Aiming for early detection. ***CMAJ : Canadian Medical Association Journal***, ***189***(12), E459–E463. https://doi.org/10.1503/cmaj.160074

Goodfellow, I., Bengio, Y., & Courville, A. (2016a). Convolutional Networks. In ***Deep learning*** (pp. 330–372). The MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016b). ***Deep learning***. The MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016c). Machine Learning Basics. In ***Deep learning*** (pp. 98–165). The MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. ***Advances in Neural Information Processing Systems***, ***27***.

Grabovska, Y., Mackay, A., O'Hare, P., Crosier, S., Finetti, M., Schwalbe, E. C., Pickles, J. C., Fairchild, A. R., Avery, A., Cockle, J., Hill, R., Lindsey, J., Hicks, D., Kristiansen, M., Chalker, J., Anderson, J., Hargrave, D., Jacques, T. S., Straathof, K., ... Williamson, D. (2020). Pediatric pan-central nervous system tumor analysis of immune-cell infiltration identifies correlates of antitumor immunity. ***Nature Communications***, ***11***(1), 4324. https://doi.org/10.1038/s41467-020-18070-y

Graf, R., Schmitt, J., Schlaeger, S., Möller, H. K., Sideri-Lampretsa, V., Sekuboyina, A., Krieg, S. M., Wiestler, B., Menze, B., Rueckert, D., & Kirschke, J. S. (2023). Denoising diffusion-based MRI to CT image translation enables automated spinal segmentation. ***European Radiology Experimental***, ***7***(1), 1–14. https://doi.org/10.1186/s41747-023-00385-2

Graham, M. S., Pinaya, W. H. L., Wright, P., Tudosiu, P.-D., Mah, Y. H., Teo, J. T., Jäger, H. R., Werring, D., Nachev, P., Ourselin, S., & Cardoso, M. J. (2023). Unsupervised 3D Out-of-Distribution Detection with Latent Diffusion Models. In ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2023*** (pp. 446–456, Vol. 14220). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43907-0_43

Grahl, S., Pongratz, V., Schmidt, P., Engl, C., Bussas, M., Radetz, A., Gonzalez-Escamilla, G., Groppa, S., Zipp, F., Lukas, C., Kirschke, J., Zimmer, C., Hoshi, M., Berthele, A., Hemmer, B., & Mühlau, M. (2019). Evidence for a white matter lesion size threshold to support the diagnosis of relapsing

remitting multiple sclerosis. ***Multiple Sclerosis and Related Disorders***, ***29***, 124–129. https://doi.org/10.1016/j.msard.2019.01.042

Gröbner, S. N., Worst, B. C., Weischenfeldt, J., Buchhalter, I., Kleinheinz, K., Rudneva, V. A., Johann, P. D., Balasubramanian, G. P., Segura-Wang, M., Brabetz, S., Bender, S., Hutter, B., Sturm, D., Pfaff, E., Hübschmann, D., Zipprich, G., Heinold, M., Eils, J., Lawerenz, C., ... Pfister, S. M. (2018). The landscape of genomic alterations across childhood cancers. ***Nature***, ***555***(7696), 321–327. https://doi.org/10.1038/nature25480

Gudbjartsson, H., & Patz, S. (1995). The Rician Distribution of Noisy MRI Data. ***Magnetic resonance in medicine***, ***34***(6), 910–914.

Guohua, C., Mengyan, L., Linyang, H., & Lingqiang, M. (2020). Multi-branch Learning Framework with Different Receptive Fields Ensemble for Brain Tumor Segmentation. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 274–284, Vol. 11993). Springer International Publishing. https://doi.org/10.1007/978-3-030-46643-5_27

Habib, A.-R., Xu, Y., Bock, K., Mohanty, S., Sederholm, T., Weeks, W. B., Dodhia, R., Ferres, J. L., Perry, C., Sacks, R., & Singh, N. (2023). Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy. ***Scientific Reports***, ***13***(1), 5368. https://doi.org/10.1038/s41598-023-31921-0

Han, X., Xie, Z., Chen, Q., Li, X., & Yang, H. (2023). Learning the degradation distribution for medical image superresolution via sparse swin transformer. ***Computers & Graphics***, ***114***, 168–178. https://doi.org/10.1016/j.cag.2023.06.003

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries***, 272–284. https://doi.org/10.1007/978-3-031-08999-2_22

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2022). UNETR: Transformers for 3D Medical Image Segmentation, 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. ***2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***, 770–778. https://doi.org/10.1109/CVPR.2016.90

Henson, J. W., Ulmer, S., & Harris, G. J. (2008). Brain Tumor Imaging in Clinical Trials. ***American Journal of Neuroradiology***, ***29***(3), 419–424. https://doi.org/10.3174/ajnr.A0963

Herington, J., McCradden, M. D., Creel, K., Boellaard, R., Jones, E. C., Jha, A. K., Rahmim, A., Scott, P. J., Sunderland, J. J., Wahl, R. L., Zuehlsdorff, S., & Saboury, B. (2023). Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation. ***Journal of Nuclear Medicine***, ***64***(12), 1848–1854. https://doi.org/10.2967/jnumed.123.266080

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross-attention control. ***The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023***.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. ***Advances in Neural Information Processing Systems***, ***30***.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. ***Advances in Neural Information Processing Systems***, ***33***, 6840–6851.

Ho, J., & Salimans, T. (2021). Classifier-free diffusion guidance. ***NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications***.

Hogea, C., Biros, G., Abraham, F., & Davatzikos, C. (2007). A robust framework for soft tissue simulations with application to modeling brain tumor mass

effect in 3D MR images. ***Physics in Medicine & Biology***, ***52***(23), 6893. https://doi.org/10.1088/0031-9155/52/23/008

Hollis, K. F. (2016). To Share or Not to Share: Ethical Acquisition and Use of Medical Data. ***AMIA Summits on Translational Science Proceedings***, ***2016***, 420.

Horé, A., & Ziou, D. (2010). Image Quality Metrics: PSNR vs. SSIM. ***2010 20th International Conference on Pattern Recognition***, 2366–2369. https://doi.org/10.1109/ICPR.2010.579

Hossain, J., Xiao, W., Tayeb, M., & Khan, S. (2021). Epidemiology and prognostic factors of pediatric brain tumor survival in the US: Evidence from four decades of population data. ***Cancer epidemiology***, ***72***, 101942. https://doi.org/10.1016/j.canep.2021.101942

Hu, Z., Li, L., Sui, A., Wu, G., Wang, Y., & Yu, J. (2023). An efficient R-Transformer network with dual encoders for brain glioma segmentation in MR images. ***Biomedical Signal Processing and Control***, ***79***. https://doi.org/10.1016/j.bspc.2022.104034

Huang, B., Xiao, H., Liu, W., Zhang, Y., Wu, H., Wang, W., Yang, Y., Yang, Y., Miller, G. W., Li, T., & Cai, J. (2021). MRI Super-Resolution via Realistic Downsampling with Adversarial Learning. ***Physics in medicine and biology***, ***66***(20), 10.1088/1361–6560/ac232e. https://doi.org/10.1088/1361-6560/ac232e

Huang, J., Shlobin, N. A., Lam, S. K., & DeCuypere, M. (2022). Artificial Intelligence Applications in Pediatric Brain Tumor Imaging: A Systematic Review. ***World Neurosurgery***, ***157***, 99–105. https://doi.org/10.1016/j.wneu.2021.10.068

Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. ***Nature Methods***, ***18***(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

IXI Dataset. (2004–2006). ***Information eXtraction from images (IXI) dataset***.

J, B. R., Sood, A., Pattnaik, T., Malhotra, R., Nayyar, V., Narayan, B., Mishra, D., & Surya, V. (2025). Medical imaging privacy: A systematic scoping review of key parameters in dataset construction and data protection. ***Journal of Medical Imaging and Radiation Sciences***, ***56***(5), 101914. https://doi.org/10.1016/j.jmir.2025.101914

Jadon, S. (2020). A survey of loss functions for semantic segmentation. ***2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)***, 1–7. https://doi.org/10.1109/CIBCB48159.2020.9277638

Jaju, A., Yeom, K. W., & Ryan, M. E. (2022). MR Imaging of Pediatric Brain Tumors. ***Diagnostics***, ***12***(4), 961. https://doi.org/10.3390/diagnostics12040961

Jiang, H., & Nachum, O. (2019). ***Identifying and Correcting Label Bias in Machine Learning***.

Jiang, Z., Ding, C., Liu, M., & Tao, D. (2020). Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 231–241, Vol. 11992). Springer International Publishing. https://doi.org/10.1007/978-3-030-46640-4_22

Jung, E., Luna, M., & Park, S. H. (2021). Conditional GAN with an Attention-Based Generator and a 3D Discriminator for 3D Medical Image Generation. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2021***, 318–328. https://doi.org/10.1007/978-3-030-87231-1_31

Kabasawa, H. (2021). MR Imaging in the 21st Century: Technical Innovation over the First Two Decades. ***Magnetic Resonance in Medical Sciences***, ***21***(1), 71–82. https://doi.org/10.2463/mrms.rev.2021-0011

Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., Saleh, A., Makowski, M., Rueckert, D., & Braren, R. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. ***Nature Machine Intelligence***, ***3***(6), 473–484. https://doi.org/10.1038/s42256-021-00337-8

Kalantar, R., Lin, G., Winfield, J. M., Messiou, C., Koh, D.-M., & Blackledge, M. D. (2023). MED-INPAINT: Medical Image Synthesis Using Multi-Level Conditional Inpainting with a Denoising Diffusion Probabilistic Model and Adaptive Contrast Priors. ***2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)***, 403–413. https://doi.org/10.1109/MedAI59581.2023.00061

Kao, P.-Y., Shailja, S., Jiang, J., Zhang, A., & Khan, A. (2020). Improving Patch-Based Convolutional Neural Networks for MRI Brain Tumor Segmentation by Leveraging Location Information. ***Frontiers in Neuroscience***, ***13***. https://doi.org/10.3389/fnins.2019.01449

Karimi, D., & Salcudean, S. E. (2020). Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks. ***IEEE Transactions on Medical Imaging***, ***39***(2), 499–513. https://doi.org/10.1109/TMI.2019.2930068

Kaur, B., Lemaître, P., Mehta, R., Sepahvand, N. M., Precup, D., Arnold, D., & Arbel, T. (2019). Improving Pathological Structure Segmentation via Transfer Learning Across Diseases. ***Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data***, 90–98. https://doi.org/10.1007/978-3-030-33391-1_11

Kayal, S., Chen, S., & de Bruijne, M. (2020). Region-of-Interest Guided Supervoxel Inpainting for Self-supervision. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2020***, 500–509. https://doi.org/10.1007/978-3-030-59710-8_49

Kazerooni, A. F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S. M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., Bagheri, S., Baid, U., Bergquist, T., Borja, A. J., Calabrese, E., Chung, V., Conte, G.-M., Dako, F., Eddy, J., ... Linguraru, M. G. (2024). The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). ***ArXiv***, arXiv:2305.17033v7.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. ***Medical Image Analysis***, ***88***, 102846. https://doi.org/10.1016/j.media.2023.102846

Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J. N., & Truhn, D. (2023). Denoising diffusion probabilistic models for 3D medical image generation. ***Scientific Reports***, ***13***(1), 7303. https://doi.org/10.1038/s41598-023-34341-2

Kim, K., Na, Y., Ye, S.-J., Lee, J., Ahn, S. S., Park, J. E., & Kim, H. (2024). Controllable Text-to-Image Synthesis for Multi-Modality MR Images, 7921–7930. https://doi.org/10.1109/WACV57701.2024.00775

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. ***2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings***.

Kline, C., Forester, C., & Banerjee, A. (2017). Ependymoma. In ***Pediatric CNS Tumors*** (pp. 69–92). Springer International Publishing. https://doi.org/10.1007/978-3-319-30789-3

Knoppers, B. M., & Thorogood, A. M. (2017). Ethics and Big Data in health. ***Current Opinion in Systems Biology***, ***4***, 53–57. https://doi.org/10.1016/j.coisb.2017.07.001

Konz, N., Chen, Y., Dong, H., & Mazurowski, M. A. (2024). Anatomically-Controllable Medical Image Generation with Segmentation-Guided Diffusion Models. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2024***, 88–98. https://doi.org/10.1007/978-3-031-72104-5_9

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. ***Advances in Neural Information Processing Systems***, ***25***.

Kush, R. D., Warzel, D., Kush, M. A., Sherman, A., Navarro, E. A., Fitzmartin, R., Pétavy, F., Galvez, J., Becnel, L. B., Zhou, F. L., Harmon, N., Jauregui, B., Jackson, T., & Hudson, L. (2020). FAIR data sharing: The roles of common data elements and harmonization. ***Journal of Biomedical Informatics***, ***107***, 103421. https://doi.org/10.1016/j.jbi.2020.103421

Kwak, H.-S., Hwang, S., Chung, G.-H., Song, J.-S., & Choi, E.-J. (2015). Detection of small brain metastases at 3 T: Comparing the diagnostic performances of contrast-enhanced T1-weighted SPACE, MPRAGE, and 2D FLASH imaging. ***Clinical Imaging***, ***39***(4), 571–575. https://doi.org/10.1016/j.clinimag.2015.02.010

Lanphear, J., & Sarnaik, S. (2014). Presenting Symptoms of Pediatric Brain Tumors Diagnosed in the Emergency Department. ***Pediatric Emergency Care***, ***30***(2), 77–80. https://doi.org/10.1097/PEC.0000000000000074

Larobina, M., & Murino, L. (2014). Medical Image File Formats. ***Journal of Digital Imaging***, ***27***(2), 200–206. https://doi.org/10.1007/s10278-013-9657-9

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., & Langlotz, C. P. (2020). Ethics of Using and Sharing Clinical Imaging Data for Artificial Intelligence: A Proposed Framework. ***Radiology***, ***295***(3), 675–682. https://doi.org/10.1148/radiol.2020192536

Lassaletta, A., Bouffet, E., Mabbott, D., & Kulkarni, A. V. (2015). Functional and neuropsychological late outcomes in posterior fossa tumors in children. ***Child's Nervous System: ChNS: Official Journal of the International Society for Pediatric Neurosurgery***, ***31***(10), 1877–1890. https://doi.org/10.1007/s00381-015-2829-9

Lee, C. S., Nagy, P. G., Weaver, S. J., & Newman-Toker, D. E. (2013). Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. ***American***

Journal of Roentgenology, ***201***(3), 611–617. https://doi.org/10.2214/AJR.12.10375

Lepcha, D. C., Goyal, B., Dogra, A., & Goyal, V. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. ***Information Fusion***, ***91***, 230–260. https://doi.org/10.1016/j.inffus.2022.10.007

Li, L., Wei, M., Liu, B., Atchaneeyasakul, K., Zhou, F., Pan, Z., Kumar, S. A., Zhang, J. Y., Pu, Y., Liebeskind, D. S., & Scalzo, F. (2021). Deep Learning for Hemorrhagic Lesion Detection and Segmentation on Brain CT Images. ***IEEE Journal of Biomedical and Health Informatics***, ***25***(5), 1646–1659. https://doi.org/10.1109/JBHI.2020.3028243

Li, Q., Yu, Z., Wang, Y., & Zheng, H. (2020). TumorGAN: A Multi-Modal Data Augmentation Framework for Brain Tumor Segmentation. ***Sensors***, ***20***(15), 4203. https://doi.org/10.3390/s20154203

Li, X., Shang, K., Wang, G., & Butala, M. D. (2023). ***DDMM-Synth: A Denoising Diffusion Model for Cross-modal Medical Image Synthesis with Sparse-view Measurement Embedding***. arXiv: 2303.15770 [eess]. https://doi.org/10.48550/arXiv.2303.15770

Liang, J., Yang, C., Zhong, J., & Ye, X. (2022). BTSwin-Unet: 3D U-shaped Symmetrical Swin Transformer-based Network for Brain Tumor Segmentation with Self-supervised Pre-training. ***Neural Processing Letters***. https://doi.org/10.1007/s11063-022-10919-1

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal Loss for Dense Object Detection. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, ***42***(2), 318–327. https://doi.org/10.1109/TPAMI.2018.2858826

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In ***Computer Vision - ECCV 2014*** (pp. 740–755, Vol. 8693). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Lipková, J., Menze, B., Wiestler, B., Koumoutsakos, P., & Lowengrub, J. S. (2022). Modelling glioma progression, mass effect and intracranial pressure in patient anatomy. ***Journal of The Royal Society Interface***, ***19***(188), 20210922. https://doi.org/10.1098/rsif.2021.0922

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. https://doi.org/10.1016/j.media.2017.07.005

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ***2021 IEEE/CVF International Conference on Computer Vision (ICCV)***, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

Llambias, S. N., Machnio, J., Munk, A., Ambsdorf, J., Nielsen, M., & Ghazi, M. M. (2024). ***Yucca: A Deep Learning Framework For Medical Image Analysis***. arXiv: 2407.19888 [cs]. https://doi.org/10.48550/arXiv.2407.19888

Loesch, N., Catchpoole, D. R., & Kennedy, P. J. (2025). Three-Dimensional Latent Diffusion Model for Weakly-Supervised Brain Tumour Segmentation. In ***Artificial Intelligence in Medicine*** (pp. 242–251, Vol. 15734). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-95838-0_24

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: A summary. ***Neuro-Oncology***, ***23***(8), 1231–1251. https://doi.org/10.1093/neuonc/noab106

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. ***Zeitschrift für Medizinische Physik***, ***29***(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

Luo, L., Li, Y., Chai, Z., Lin, H., Heng, P.-A., & Chen, H. (2024). Scale-Aware Super-Resolution Network With Dual Affinity Learning for Lesion Segmentation From Medical Images. ***IEEE Transactions on Neural Networks and Learning Systems***, 1–14. https://doi.org/10.1109/TNNLS.2024.3477947

Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., Zhou, X., Li, Y., Rusch, M. C., Easton, J., Huether, R., Gonzalez-Pena, V., Wilkinson, M. R., Hermida, L. C., Davis, S., Sioson, E., Pounds, S., Cao, X., Ries, R. E., … Zhang, J. (2018). Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. ***Nature***, ***555***(7696), 371–376. https://doi.org/10.1038/nature25795

Macdonald, D. R., Cascino, T. L., Schold, S. C., & Cairncross, J. G. (1990). Response criteria for phase II studies of supratentorial malignant glioma. ***Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology***, ***8***(7), 1277–1280. https://doi.org/10.1200/JCO.1990.8.7.1277

Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., … Jäger, P. F. (2024). Metrics reloaded: Recommendations for image analysis validation. ***Nature Methods***, ***21***(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

Mamlouk, M. D., Bryant, S. O., Cha, S., & Barkovich, A. J. (2017). Modern Neuroimaging of Pediatric Brain Tumors. In ***Pediatric CNS Tumors*** (pp. 273–300). Springer International Publishing. https://doi.org/10.1007/978-3-319-30789-3_13

Mao, Y., Jiang, L., Chen, X., & Li, C. (2023). DisC-Diff: Disentangled Conditional Diffusion Model for Multi-contrast MRI Super-Resolution. In ***Medical Image***

Computing and Computer Assisted Intervention - MICCAI 2023 (pp. 387–397, Vol. 14229). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43999-5_37

Maqsood, M., Nazir, F., Khan, U., Aadil, F., Jamal, H., Mehmood, I., & Song, O.-y. (2019). Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans. ***Sensors (Basel, Switzerland)***, ***19***(11), 2645. https://doi.org/10.3390/s19112645

Martucci, M., Russo, R., Schimperna, F., D'Apolito, G., Panfili, M., Grimaldi, A., Perna, A., Ferranti, A. M., Varcasia, G., Giordano, C., & Gaudino, S. (2023). Magnetic Resonance Imaging of Primary Adult Brain Tumors: State of the Art and Future Perspectives. ***Biomedicines***, ***11***(2), 364. https://doi.org/10.3390/biomedicines11020364

Matsumoto, T., Miura, T., & Yanai, N. (2023). Membership Inference Attacks against Diffusion Models. ***2023 IEEE Security and Privacy Workshops (SPW)***, 77–83. https://doi.org/10.1109/SPW59333.2023.00013

Mehta, R., Filos, A., Baid, U., Sako, C., McKinley, R., Rebsamen, M., Dätwyler, K., Meier, R., Radojewski, P., Murugesan, G. K., Nalawade, S., Ganesh, C., Wagner, B., Yu, F. F., Fei, B., Madhuranthakam, A. J., Maldjian, J. A., Daza, L., Gómez, C., … Arbel, T. (2021). ***QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation – Analysis of Ranking Metrics and Benchmarking Results***.

Meissen, F., Kaissis, G., & Rueckert, D. (2022). Challenging Current Semi-supervised Anomaly Segmentation Methods for Brain MRI. ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries***, 63–74. https://doi.org/10.1007/978-3-031-08999-2_5

Meng, X., Sun, K., Xu, J., He, X., & Shen, D. (2024). Multi-Modal Modality-Masked Diffusion Network for Brain MRI Synthesis With Random Modality Missing. ***IEEE Transactions on Medical Imaging***, ***43***(7), 2587–2598. https://doi.org/10.1109/TMI.2024.3368664

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). ***IEEE Transactions on Medical Imaging***, ***34***(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

Miller, K. D., Ostrom, Q. T., Kruchko, C., Patil, N., Tihan, T., Cioffi, G., Fuchs, H. E., Waite, K. A., Jemal, A., Siegel, R. L., & Barnholtz-Sloan, J. S. (2021). Brain and other central nervous system tumor statistics, 2021. ***CA: A Cancer Journal for Clinicians***, ***71***(5), 381–406. https://doi.org/10.3322/caac.21693

Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image Segmentation Using Deep Learning: A Survey. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 1–1. https://doi.org/10.1109/TPAMI.2021.3059968

Moawad, A. W., Janas, A., Baid, U., Ramakrishnan, D., Saluja, R., Ashraf, N., Jekel, L., Amiruddin, R., Adewole, M., Albrecht, J., Anazodo, U., Aneja, S., Anwar, S. M., Bergquist, T., Calabrese, E., Chiang, V., Chung, V., Conte, G. M. M., Dako, F., ... Aboian, M. (2024). ***The Brain Tumor Segmentation (BraTS-METS) Challenge 2023: Brain Metastasis Segmentation on Pre-treatment MRI***. arXiv: 2306.00838 [eess, q-bio]. https://doi.org/10.48550/arXiv.2306.00838

Moser, B. B., Shanbhag, A. S., Raue, F., Frolov, S., Palacio, S., & Dengel, A. (2025). Diffusion Models, Image Super-Resolution, and Everything: A Survey. ***IEEE Transactions on Neural Networks and Learning Systems***, ***36***(7), 11793–11813. https://doi.org/10.1109/TNNLS.2024.3476671

Mueller, S., & Chang, S. (2009). Pediatric brain tumors: Current treatment strategies and future therapeutic approaches. ***Neurotherapeutics***, ***6***(3), 570–586. https://doi.org/10.1016/j.nurt.2009.04.006

Müller, S., Weickert, J., & Graf, N. (2020). Robustness of brain tumor segmentation. ***Journal of Medical Imaging***, ***7***(06). https://doi.org/10.1117/1.JMI.7.6.064006

Müller-Franzes, G., Niehues, J. M., Khader, F., Arasteh, S. T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J. N., & Truhn, D. (2023). A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. ***Scientific Reports***, ***13***(1), 12098. https://doi.org/10.1038/s41598-023-39278-0

Mulvany, T., Griffiths-King, D., Novak, J., & Rose, H. (2024). Segmentation of pediatric brain tumors using a radiologically informed, deep learning cascade. ***CoRR***, ***abs/2410.14020***. https://doi.org/10.48550/ARXIV.2410.14020

Myronenko, A. (2019). 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 311–320, Vol. 11384). Springer International Publishing. https://doi.org/10.1007/978-3-030-11726-9_28

Nair, T., Precup, D., Arnold, D. L., & Arbel, T. (2020). Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. ***Medical Image Analysis***, ***59***, 101557. https://doi.org/10.1016/j.media.2019.101557

Najjar, R. (2023). Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. ***Diagnostics***, ***13***(17), 2760. https://doi.org/10.3390/diagnostics13172760

Nalepa, J., Marcinkiewicz, M., & Kawulok, M. (2019). Data Augmentation for Brain-Tumor Segmentation: A Review. ***Frontiers in Computational Neuroscience***, ***13***, 83. https://doi.org/10.3389/fncom.2019.00083

Nichol, A. Q., & Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. ***Proceedings of the 38th International Conference on Machine Learning***, 8162–8171.

Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. ***IEEE Transactions on Systems, Man, and Cybernetics***, ***9***(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., & Maier, A. (2022). Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. ***Scientific Reports***, ***12***(1), 14851. https://doi.org/10.1038/s41598-022-19045-3

Painuli, D., Bhardwaj, S., & köse, U. (2022). Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review. ***Computers in Biology and Medicine***, ***146***, 105580. https://doi.org/10.1016/j.compbiomed.2022.105580

Palmer, S. L., Goloubeva, O., Reddick, W. E., Glass, J. O., Gajjar, A., Kun, L., Merchant, T. E., & Mulhern, R. K. (2001). Patterns of intellectual development among survivors of pediatric medulloblastoma: A longitudinal analysis. ***Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology***, ***19***(8), 2302–2308. https://doi.org/10.1200/JCO.2001.19.8.2302

Park, T., Liu, M.-Y., Wang, T.-C., & Zhu, J.-Y. (2019). Semantic Image Synthesis With Spatially-Adaptive Normalization. ***2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 2332–2341. https://doi.org/10.1109/CVPR.2019.00244

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. ***Advances in Neural Information Processing Systems***, ***32***.

Pati, S., Singh, A., Rathore, S., Gastounioti, A., Bergman, M., Ngo, P., Ha, S. M., Bounias, D., Minock, J., Murphy, G., Li, H., Bhattarai, A., Wolf, A., Sridaran, P., Kalarot, R., Akbari, H., Sotiras, A., Thakur, S. P., Verma, R., … Bakas, S. (2020). The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview.

In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 380–394, Vol. 11993). Springer International Publishing. https://doi.org/10.1007/978-3-030-46643-5_38

Patil, S. S., Rajak, R., Ramteke, M., & Rathore, A. S. (2025). MMIT-DDPM - Multilateral medical image translation with class and structure supervised diffusion-based model. ***Computers in Biology and Medicine***, ***185***, 109501. https://doi.org/10.1016/j.compbiomed.2024.109501

Patrinos, D., Knoppers, B. M., Laplante, D. P., Rahbari, N., & Wazana, A. (2022). Sharing and Safeguarding Pediatric Data. ***Frontiers in Genetics***, ***13***, 872586. https://doi.org/10.3389/fgene.2022.872586

Peng, J., & Wang, Y. (2021). Medical Image Segmentation With Limited Supervision: A Review of Deep Network Models. ***IEEE Access***, ***9***, 36827–36851. https://doi.org/10.1109/ACCESS.2021.3062380

Peng, J., Estrada, G., Pedersoli, M., & Desrosiers, C. (2020). Deep co-training for semi-supervised image segmentation. ***Pattern Recognition***, ***107***, 107269. https://doi.org/10.1016/j.patcog.2020.107269

Peng, W., Adeli, E., Bosschieter, T., Park, S. H., Zhao, Q., & Pohl, K. M. (2023). Generating Realistic Brain MRIs via a Conditional Diffusion Probabilistic Model. In ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2023*** (pp. 14–24, Vol. 14227). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43993-3_2

Pérez-García, F., Sparks, R., & Ourselin, S. (2021). TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. ***Computer Methods and Programs in Biomedicine***, ***208***, 106236. https://doi.org/10.1016/j.cmpb.2021.106236

Pfister, S. M., Reyes-Múgica, M., Chan, J. K. C., Hasle, H., Lazar, A. J., Rossi, S., Ferrari, A., Jarzembowski, J. A., Pritchard-Jones, K., Hill, D. A., Jacques, T. S., Wesseling, P., López Terrada, D. H., von Deimling, A., Kratz, C. P., Cree, I. A., & Alaggio, R. (2022). A Summary of the Inaugural WHO Classification of Pediatric Tumors: Transitioning from the Optical into the Molecular Era.

Cancer Discovery, ***12***(2), 331–355. https://doi.org/10.1158/2159-8290.CD-21-1094

Pinaya, W. H. L., 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., Werring, D., Rees, G., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models. In ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*** (pp. 705–714, Vol. 13438). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16452-1_67

Pinaya, W. H. L., Tudosiu, P.-D., Gray, R., Rees, G., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. ***Medical Image Analysis***, ***79***, 102475. https://doi.org/10.1016/j.media.2022.102475

Prasanna, P., Mitra, J., Beig, N., Nayate, A., Patel, J., Ghose, S., Thawani, R., Partovi, S., Madabhushi, A., & Tiwari, P. (2019). Mass Effect Deformation Heterogeneity (MEDH) on Gadolinium-contrast T1-weighted MRI is associated with decreased survival in patients with right cerebral hemisphere Glioblastoma: A feasibility study. ***Scientific Reports***, ***9***, 1145. https://doi.org/10.1038/s41598-018-37615-2

Price, W. N., & Cohen, I. G. (2019). Privacy in the Age of Medical Big Data. ***Nature medicine***, ***25***(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Pui, C.-H., Gajjar, A. J., Kane, J. R., Qaddoumi, I. A., & Pappo, A. S. (2011). Challenging issues in pediatric oncology. ***Nature Reviews. Clinical Oncology***, ***8***(9), 540–549. https://doi.org/10.1038/nrclinonc.2011.95

Rabe, M. N., & Staats, C. (2022). ***Self-attention Does Not Need $O(n^2)$ Memory***. arXiv: 2112.05682 [cs]. https://doi.org/10.48550/arXiv.2112.05682

Renieblas, G. P., Nogués, A. T., González, A. M., Gómez-Leon, N., & Del Castillo, E. G. (2017). Structural similarity index family for image quality assessment in

radiological images. ***Journal of Medical Imaging (Bellingham, Wash.)***, ***4***(3), 035501. https://doi.org/10.1117/1.JMI.4.3.035501

Ris, M. D., Packer, R., Goldwein, J., Jones-Wallace, D., & Boyett, J. M. (2001). Intellectual outcome after reduced-dose radiation therapy plus adjuvant chemotherapy for medulloblastoma: A Children's Cancer Group study. ***Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology***, ***19***(15), 3470–3476. https://doi.org/10.1200/JCO.2001.19.15.3470

Rohlfing, T., Zahr, N. M., Sullivan, E. V., & Pfefferbaum, A. (2010). The SRI24 multichannel atlas of normal adult human brain structure. ***Human Brain Mapping***, ***31***(5), 798–819. https://doi.org/10.1002/hbm.20906

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. ***2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

Ronneberger, O., Fischer, P., & Brox, T. (2015). ***U-Net: Convolutional Networks for Biomedical Image Segmentation***.

Rosas González, S., Birgui Sekou, T., Hidane, M., & Tauber, C. (2020). 3D Automatic Brain Tumor Segmentation Using a Multiscale Input U-Net Network. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 113–123, Vol. 11993). Springer International Publishing. https://doi.org/10.1007/978-3-030-46643-5_11

Sabeghi, P., Zarand, P., Zargham, S., Golestany, B., Shariat, A., Chang, M., Yang, E., Rajagopalan, P., Phung, D. C., & Gholamrezanezhad, A. (2024). Advances in Neuro-Oncological Imaging: An Update on Diagnostic Approach to Brain Tumors. ***Cancers***, ***16***(3), 576. https://doi.org/10.3390/cancers16030576

Sadighi, Z. S., Curtis, E., Zabrowksi, J., Billups, C., Gajjar, A., Khan, R., & Qaddoumi, I. (2018). Neurologic impairments from pediatric low-grade glioma by tumor location and timing of diagnosis. ***Pediatric Blood & Cancer***, ***65***(8), e27063. https://doi.org/10.1002/pbc.27063

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2023). Image Super-Resolution via Iterative Refinement. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, ***45***(4), 4713–4726. https://doi.org/10.1109/TPAMI.2022.3204461

Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In ***Machine Learning in Medical Imaging*** (pp. 379–387, Vol. 10541). Springer International Publishing. https://doi.org/10.1007/978-3-319-67389-9_44

Sanchez, P., Kascenas, A., Liu, X., O'Neil, A. Q., & Tsaftaris, S. A. (2022). What is Healthy? Generative Counterfactual Diffusion for Lesion Localization. In ***Deep Generative Models*** (pp. 34–44, Vol. 13609). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-18576-2_4

Savelli, B., Bria, A., Molinara, M., Marrocco, C., & Tortorella, F. (2020). A multi-context CNN ensemble for small lesion detection. ***Artificial Intelligence in Medicine***, ***103***, 101749. https://doi.org/10.1016/j.artmed.2019.101749

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., & Schmidt-Erfurth, U. (2019). F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. ***Medical Image Analysis***, ***54***, 30–44. https://doi.org/10.1016/j.media.2019.01.010

Shaari, H., Kevrić, J., Jukić, S., Bešić, L., Jokić, D., Ahmed, N., & Rajs, V. (2021). Deep Learning-Based Studies on Pediatric Brain Tumors Imaging: Narrative Review of Techniques and Challenges. ***Brain Sciences***, ***11***(6), 716. https://doi.org/10.3390/brainsci11060716

Shao, D., Qin, L., Xiang, Y., Ma, L., & Xu, H. (2023). Medical image blind super-resolution based on improved degradation process. ***IET Image Processing***, ***17***(5), 1615–1625. https://doi.org/10.1049/ipr2.12742

Shaw, R., Sudre, C., Ourselin, S., & Cardoso, M. J. (2019). MRI k-Space Motion Artefact Augmentation: Model Robustness and Task-Specific Uncertainty.

Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, 427–436.

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. ***Annual Review of Biomedical Engineering***, ***19***(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Shin, M., Seo, M., Lee, K., & Yoon, K. (2024). Super-resolution techniques for biomedical applications and challenges. ***Biomedical Engineering Letters***. https://doi.org/10.1007/s13534-024-00365-4

Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. ***CA: A Cancer Journal for Clinicians***, ***74***(1), 12–49. https://doi.org/10.3322/caac.21820

Siegel, R. L., Miller, K. D., & Jemal, A. (2019). Cancer statistics, 2019. ***CA: A Cancer Journal for Clinicians***, ***69***(1), 7–34. https://doi.org/10.3322/caac.21551

Singh, S. P., Wang, L., Gupta, S., Goli, H., Padmanabhan, P., & Gulyás, B. (2020). 3D Deep Learning on Medical Images: A Review. ***Sensors (Basel, Switzerland)***, ***20***(18), 5097. https://doi.org/10.3390/s20185097

Sizikova, E., Badal, A., Delfino, J. G., Lago, M., Nelson, B., Saharkhiz, N., Sahiner, B., Zamzmi, G., & Badano, A. (2024). Synthetic data in radiological imaging: Current state and future outlook. ***BJR—Artificial Intelligence***, ***1***(1), ubae007. https://doi.org/10.1093/bjrai/ubae007

Sled, J., Zijdenbos, A., & Evans, A. (1998). A nonparametric method for automatic correction of intensity nonuniformity in MRI data. ***IEEE Transactions on Medical Imaging***, ***17***(1), 87–97. https://doi.org/10.1109/42.668698

Smith, J. S., Chang, E. F., Lamborn, K. R., Chang, S. M., Prados, M. D., Cha, S., Tihan, T., Vandenberg, S., McDermott, M. W., & Berger, M. S. (2008). Role of extent of resection in the long-term outcome of low-grade hemispheric gliomas. ***Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology***, ***26***(8), 1338–1345. https://doi.org/10.1200/JCO.2007.13.9337

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. ***Proceedings of the 32nd International Conference on Machine Learning***, ***37***, 2256–2265.

Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. ***9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021***.

Song, Y., & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In ***Proceedings of the 33rd International Conference on Neural Information Processing Systems*** (pp. 11918–11930). Curran Associates Inc.

Soomro, T. A., Zheng, L., Afifi, A. J., Ali, A., Soomro, S., Yin, M., & Gao, J. (2023). Image Segmentation for MR Brain Tumor Detection Using Machine Learning: A Review. ***IEEE Reviews in Biomedical Engineering***, ***16***, 70–90. https://doi.org/10.1109/RBME.2022.3185292

Sørensen, P. J., Carlsen, J. F., Larsen, V. A., Andersen, F. L., Ladefoged, C. N., Nielsen, M. B., Poulsen, H. S., & Hansen, A. E. (2023). Evaluation of the HD-GLIO Deep Learning Algorithm for Brain Tumour Segmentation on Postoperative MRI. ***Diagnostics***, ***13***(3), 363. https://doi.org/10.3390/diagnostics13030363

Sterzing, F., Engenhart-Cabillic, R., Flentje, M., & Debus, J. (2011). Image-Guided Radiotherapy. ***Deutsches Ärzteblatt International***, ***108***(16), 274–280. https://doi.org/10.3238/arztebl.2011.0274

Storey, P., Chen, Q., Li, W., Edelman, R. R., & Prasad, P. V. (2002). Band artifacts due to bulk motion. ***Magnetic Resonance in Medicine***, ***48***(6), 1028–1036. https://doi.org/10.1002/mrm.10314

Sturm, D., Pfister, S. M., & Jones, D. T. (2017). Pediatric Gliomas: Current Concepts on Diagnosis, Biology, and Clinical Management. ***Journal of Clinical Oncology***, ***35***(21), 2370–2377. https://doi.org/10.1200/JCO.2017.73.0242

Subramanian, S., Ghafouri, A., Scheufele, K. M., Himthani, N., Davatzikos, C., & Biros, G. (2023). Ensemble Inversion for Brain Tumor Growth Models With Mass Effect. ***IEEE transactions on medical imaging***, ***42***(4), 982–995. https://doi.org/10.1109/TMI.2022.3221913

Sudre, C. H., Cardoso, M. J., & Ourselin, S. (2017). Longitudinal segmentation of age-related white matter hyperintensities. ***Medical Image Analysis***, ***38***, 50–64. https://doi.org/10.1016/j.media.2017.02.007

Sun, C., Shrivastava, A., Singh, S., & Gupta, A. (2017). Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. ***2017 IEEE International Conference on Computer Vision (ICCV)***, 843–852. https://doi.org/10.1109/ICCV.2017.97

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. ***BMC Medical Imaging***, ***15***(1), 29. https://doi.org/10.1186/s12880-015-0068-x

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A Survey on Deep Transfer Learning. ***Artificial Neural Networks and Machine Learning - ICANN 2018***, 270–279. https://doi.org/10.1007/978-3-030-01424-7_27

Tao, Q., Ge, Z., Cai, J., Yin, J., & See, S. (2019). Improving Deep Lesion Detection Using 3D Contextual and Spatial Attention. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2019***, 185–193. https://doi.org/10.1007/978-3-030-32226-7_21

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and Mitigating Bias in Imaging Artificial Intelligence. ***RadioGraphics***, ***44***(5), e230067. https://doi.org/10.1148/rg.230067

Thompson, B. H., Di Caterina, G., & Voisey, J. P. (2022). Pseudo-Label Refinement Using Superpixels for Semi-Supervised Brain Tumour Segmentation. ***2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)***, 1–5. https://doi.org/10.1109/ISBI52829.2022.9761681

Thust, S. C., Heiland, S., Falini, A., Jäger, H. R., Waldman, A. D., Sundgren, P. C., Godi, C., Katsaros, V. K., Ramos, A., Bargallo, N., Vernooij, M. W., Yousry,

T., Bendszus, M., & Smits, M. (2018). Glioma imaging in Europe: A survey of 220 centres and recommendations for best clinical practice. ***European Radiology***, ***28***(8), 3306–3317. https://doi.org/10.1007/s00330-018-5314-5

Tokuoka, Y., Suzuki, S., & Sugawara, Y. (2019). An Inductive Transfer Learning Approach using Cycle-consistent Adversarial Domain Adaptation with Application to Brain Tumor Segmentation. ***Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering***, 44–48. https://doi.org/10.1145/3375923.3375948

Vagvala, S., Guenette, J. P., Jaimes, C., & Huang, R. Y. (2022). Imaging diagnosis and treatment selection for brain tumors in the era of molecular therapeutics. ***Cancer Imaging***, ***22***(1), 19. https://doi.org/10.1186/s40644-022-00455-5

Valverde, J. M., Imani, V., Abdollahzadeh, A., De Feo, R., Prakash, M., Ciszek, R., & Tohka, J. (2021). Transfer Learning in Magnetic Resonance Brain Imaging: A Systematic Review. ***Journal of Imaging***, ***7***(4), 66. https://doi.org/10.3390/jimaging7040066

Van Leemput, K., Maes, F., Vandermeulen, D., & Suetens, P. (1999). Automated model-based tissue classification of MR images of the brain. ***IEEE transactions on medical imaging***, ***18***(10), 897–908. https://doi.org/10.1109/42.811270

VanBerlo, B., Hoey, J., & Wong, A. (2024). A survey of the impact of self-supervised pretraining for diagnostic tasks in medical X-ray, CT, MRI, and ultrasound. ***BMC Medical Imaging***, ***24***(1), 79. https://doi.org/10.1186/s12880-024-01253-0

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: Methodological failures and recommendations for the future. ***npj Digital Medicine***, ***5***(1), 48. https://doi.org/10.1038/s41746-022-00592-y

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***.

Verburg, N., & de Witt Hamer, P. C. (2021). State-of-the-art imaging for glioma surgery. ***Neurosurgical Review***, ***44***(3), 1331–1343. https://doi.org/10.1007/s10143-020-01337-9

Villanueva-Meyer, J. E., Mabray, M. C., & Cha, S. (2017). Current Clinical Brain Tumor Imaging. ***Neurosurgery***, ***81***(3), 397–415. https://doi.org/10.1093/neuros/nyx103

Wacker, J., Ladeira, M., & Nascimento, J. E. V. (2021). Transfer Learning for Brain Tumor Segmentation. In ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*** (pp. 241–251, Vol. 12658). Springer International Publishing. https://doi.org/10.1007/978-3-030-72084-1_22

Walker, D., Wilne, S., Grundy, R., Kennedy, C., Dickson, A., Lindsell, S., Trusler, J., Dudley, J., Evans, A., Thomson, A., Lakhanpaul, M., Clough, L., Baker, M., Chu, T., Liu, J.-F., Pearson, E., Rayner, E., Thorne, E., & Franklin, S. (2016). A new clinical guideline from the Royal College of Paediatrics and Child Health with a national awareness campaign accelerates brain tumor diagnosis in UK children - "HeadSmart: Be Brain Tumour Aware". ***Neuro-Oncology***, ***18***(3), 445–454. https://doi.org/10.1093/neuonc/nov187

Wallace, B., Gokul, A., & Naik, N. (2023). EDICT: Exact Diffusion Inversion via Coupled Transformations. ***2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 22532–22541. https://doi.org/10.1109/CVPR52729.2023.02158

Wang, C., Zhao, Z., Ren, Q., Xu, Y., & Yu, Y. (2019). Dense U-net Based on Patch-Based Learning for Retinal Vessel Segmentation. ***Entropy***, ***21***(2), 168. https://doi.org/10.3390/e21020168

Wang, D., Zhang, Y., Zhang, K., & Wang, L. (2020). FocalMix: Semi-Supervised Learning for 3D Medical Image Detection, 3950–3959. https://doi.org/10.1109/CVPR42600.2020.00401

Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., & Qin, J. (2021). Boundary-Aware Transformers for Skin Lesion Segmentation. ***Medical Image Computing***

and Computer Assisted Intervention - MICCAI 2021, 206–216. https://doi.org/10.1007/978-3-030-87193-2_20

Wang, J., Yue, Z., Zhou, S., Chan, K. C. K., & Loy, C. C. (2024). Exploiting Diffusion Prior for Real-World Image Super-Resolution. ***International Journal of Computer Vision***, ***132***(12), 5929–5949. https://doi.org/10.1007/s11263-024-02168-7

Wang, J., Gao, J., Ren, J., Luan, Z., Yu, Z., Zhao, Y., & Zhao, Y. (2021). DFP-ResUNet:Convolutional Neural Network with a Dilated Convolutional Feature Pyramid for Multimodal Brain Tumor Segmentation. ***Computer Methods and Programs in Biomedicine***, ***208***, 106208. https://doi.org/10.1016/j.cmpb.2021.106208

Wang, J., Levman, J., Pinaya, W. H. L., Tudosiu, P.-D., Cardoso, M. J., & Marinescu, R. (2023). InverseSR: 3D Brain MRI Super-Resolution Using a Latent Diffusion Model. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2023***, 438–447. https://doi.org/10.1007/978-3-031-43999-5_42

Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., & Guo, Y. (2021). Unsupervised Degradation Representation Learning for Blind Super-Resolution. ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 10576–10585. https://doi.org/10.1109/CVPR46437.2021.01044

Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. ***IET Image Processing***, ***16***(5), 1243–1267. https://doi.org/10.1049/ipr2.12419

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2021***, 109–119. https://doi.org/10.1007/978-3-030-87193-2_11

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. ***2021 IEEE/CVF Inter-***

national Conference on Computer Vision Workshops (ICCVW), 1905–1914. https://doi.org/10.1109/ICCVW54120.2021.00217

Wang, Z., Simoncelli, E., & Bovik, A. (2003). Multiscale structural similarity for image quality assessment. ***The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003***, ***2***, 1398–1402 Vol.2. https://doi.org/10.1109/ACSSC.2003.1292216

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: a publication of the IEEE Signal Processing Society***, ***13***(4), 600–612. https://doi.org/10.1109/tip.2003.819861

Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., & Song, H. (2021). Unsupervised Real-world Image Super Resolution via Domain-distance Aware Training. ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 13380–13389. https://doi.org/10.1109/CVPR46437.2021.01318

Weller, M., Wick, W., Aldape, K., Brada, M., Berger, M., Pfister, S. M., Nishikawa, R., Rosenthal, M., Wen, P. Y., Stupp, R., & Reifenberger, G. (2015). Glioma. ***Nature Reviews Disease Primers***, ***1***(1), 15017. https://doi.org/10.1038/nrdp.2015.17

Wells, E. M., & Packer, R. J. (2015). Pediatric brain tumors. ***Continuum (Minneapolis, Minn.)***, ***21***, 373–396. https://doi.org/10.1212/01.CON.0000464176.96311.d1

Weninger, L., Krauhausen, I., & Merhof, D. (2019). Semantic Segmentation of Brain Tumors in MRI Data Without any Labels. ***Eurographics Workshop on Visual Computing for Biology and Medicine***, 5 pages. https://doi.org/10.2312/VCBM.20191230

Wienke, J., Dierselhuis, M. P., Tytgat, G. A. M., Künkele, A., Nierkens, S., & Molenaar, J. J. (2021). The immune landscape of neuroblastoma: Challenges and opportunities for novel therapeutic strategies in pediatric oncology. ***Eu-***

ropean Journal of Cancer (Oxford, England: 1990), ***144***, 123–150. https://doi.org/10.1016/j.ejca.2020.11.014

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. ***Radiology***, ***295***(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Wilne, S., Koller, K., Collier, J., Kennedy, C., Grundy, R., & Walker, D. (2010). The diagnosis of brain tumours in children: A guideline to assist healthcare professionals in the assessment of children who may have a brain tumour. ***Archives of Disease in Childhood***, ***95***(7), 534–539. https://doi.org/10.1136/adc.2009.162057

Wirth, F. N., Meurers, T., Johns, M., & Prasser, F. (2021). Privacy-preserving data sharing infrastructures for medical research: Systematization and comparison. ***BMC Medical Informatics and Decision Making***, ***21***(1), 242. https://doi.org/10.1186/s12911-021-01602-x

Wolleb, J., Bieder, F., Sandkühler, R., & Cattin, P. C. (2022). Diffusion Models for Medical Anomaly Detection. In ***Medical Image Computing and Computer Assisted Intervention - MICCAI 2022*** (pp. 35–45, Vol. 13438). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-16452-1_4

Wong, A., Chen, A., Wu, Y., Cicek, S., Tiard, A., Hong, B.-W., & Soatto, S. (2022). Small Lesion Segmentation in Brain MRIs with Subpixel Embedding. ***Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries***, 75–87. https://doi.org/10.1007/978-3-031-08999-2_6

Wood, M. L., & Henkelman, R. M. (1985). MR image artifacts from periodic motion. ***Medical Physics***, ***12***(2), 143–151. https://doi.org/10.1118/1.595782

Wu, H., Zhao, Z., Zhang, Y., Xie, W., & Wang, Y. (2024). MRGen: Diffusion-based controllable data engine for MRI segmentation towards unannotated modalities. ***CoRR***, ***abs/2412.04106***. https://doi.org/10.48550/ARXIV.2412.04106

Wu, S., Wu, Y., Chang, H., Su, F. T., Liao, H., Tseng, W., Liao, C., Lai, F., Hsu, F., & Xiao, F. (2021). Deep Learning-Based Segmentation of Various Brain Lesions for Radiosurgery. ***Applied Sciences***, ***11***(19), 9180. https://doi.org/10.3390/app11199180

Wu, X., Bi, L., Fulham, M., Feng, D. D., Zhou, L., & Kim, J. (2021). Unsupervised brain tumor segmentation using a symmetric-driven adversarial network. ***Neurocomputing***, ***455***, 242–254. https://doi.org/10.1016/j.neucom.2021.05.073

Wyatt, J., Leach, A., Schmon, S. M., & Willcocks, C. G. (2022). AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. ***2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)***, 649–655. https://doi.org/10.1109/CVPRW56347.2022.00080

Xu, B., Chai, Y., Galarza, C. M., Vu, C. Q., Tamrazi, B., Gaonkar, B., Macyszyn, L., Coates, T. D., Lepore, N., & Wood, J. C. (2018). Orchestral fully convolutional networks for small lesion segmentation in brain MRI. ***2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)***, 889–892. https://doi.org/10.1109/ISBI.2018.8363714

Xu, Z., Wang, Y., Lu, D., Luo, X., Yan, J., Zheng, Y., & Tong, R. K.-y. (2023). Ambiguity-selective consistency regularization for mean-teacher semi-supervised medical image segmentation. ***Medical Image Analysis***, ***88***, 102880. https://doi.org/10.1016/j.media.2023.102880

Xu, Z., Wang, Y., Lu, D., Yu, L., Yan, J., Luo, J., Ma, K., Zheng, Y., & Tong, R. K.-y. (2022). All-Around Real Label Supervision: Cyclic Prototype Consistency Learning for Semi-Supervised Medical Image Segmentation. ***IEEE Journal of Biomedical and Health Informatics***, ***26***(7), 3174–3184. https://doi.org/10.1109/JBHI.2022.3162043

Xue, Y., Xu, T., Zhang, H., Long, L. R., & Huang, X. (2018). SegAN: Adversarial Network with Multi-scale L1 Loss for Medical Image Segmentation. ***Neuroinformatics***, ***16***(3–4), 383–392. https://doi.org/10.1007/s12021-018-9377-x

Yamada, Y., Kobayashi, D., Terashima, K., Kiyotani, C., Sasaki, R., Michihata, N., Kobayashi, T., Ogiwara, H., Matsumoto, K., & Ishiguro, A. (2020). Initial symptoms and diagnostic delay in children with brain tumors at a single institution in Japan. ***Neuro-Oncology Practice***, ***8***(1), 60–67. https://doi.org/10.1093/nop/npaa062

Yang, H., Wang, Z., Liu, X., Li, C., Xin, J., & Wang, Z. (2023). Deep learning in medical image super resolution: A review. ***Applied Intelligence***, ***53***(18), 20891–20916. https://doi.org/10.1007/s10489-023-04566-9

Yeghiazaryan, V., & Voiculescu, I. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. ***Journal of Medical Imaging***, ***5***(01), 1. https://doi.org/10.1117/1.JMI.5.1.015006

Yi, X., Walia, E., & Babyn, P. (2019). Generative Adversarial Network in Medical Imaging: A Review. ***Medical Image Analysis***, ***58***, 101552. https://doi.org/10.1016/j.media.2019.101552

Yoo, Y., Ceccaldi, P., Liu, S., Re, T. J., Cao, Y., Balter, J. M., & Gibson, E. (2021). Evaluating deep learning methods in detecting and segmenting different sizes of brain metastases on 3D post-contrast T1-weighted images. ***Journal of Medical Imaging***, ***8***(3), 037001. https://doi.org/10.1117/1.JMI.8.3.037001

Yu, F., Zhang, Y., Song, S., Seff, A., & Xiao, J. (2015). LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. ***CoRR***, ***abs/1506.03365***.

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021). Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. ***Neurocomputing***, ***444***, 92–110. https://doi.org/10.1016/j.neucom.2020.04.157

Zegers, C. M. L., Posch, J., Traverso, A., Eekers, D., Postma, A. A., Backes, W., Dekker, A., & van Elmpt, W. (2021). Current applications of deep-learning in neuro-oncological MRI. ***Physica Medica***, ***83***, 161–173. https://doi.org/10.1016/j.ejmp.2021.03.003

Zhang, H., Zhang, Y., Wu, Q., Wu, J., Zhen, Z., Shi, F., Yuan, J., Wei, H., Liu, C., & Zhang, Y. (2023). Self-Supervised Arbitrary Scale Super-Resolution Framework for Anisotropic MRI. ***2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)***, 1–5. https://doi.org/10.1109/ISBI53787.2023.10230678

Zhang, K., Liang, J., Van Gool, L., & Timofte, R. (2021). Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. ***2021 IEEE/CVF International Conference on Computer Vision (ICCV)***, 4771–4780. https://doi.org/10.1109/ICCV48922.2021.00475

Zhang, L., Wen, X., Li, J.-W., Jiang, X., Yang, X.-F., & Li, M. (2023). Diagnostic error and bias in the department of radiology: A pictorial essay. ***Insights into Imaging***, ***14***(1), 1–12. https://doi.org/10.1186/s13244-023-01521-7

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. ***IEEE transactions on image processing: a publication of the IEEE Signal Processing Society***, ***20***(8), 2378–2386. https://doi.org/10.1109/TIP.2011.2109730

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding Conditional Control to Text-to-Image Diffusion Models. ***2023 IEEE/CVF International Conference on Computer Vision (ICCV)***, 3813–3824. https://doi.org/10.1109/ICCV51070.2023.00355

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. ***2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition***, 586–595. https://doi.org/10.1109/CVPR.2018.00068

Zhang, Y., Mao, Y., Lu, X., Zou, X., Huang, H., Li, X., Li, J., & Zhang, H. (2024). From single to universal: Tiny lesion detection in medical imaging. ***Artificial Intelligence Review***, ***57***(8), 1–42. https://doi.org/10.1007/s10462-024-10762-x

Zhang, Y., Jiao, R., Liao, Q., Li, D., & Zhang, J. (2023). Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation. ***Artifi-***

cial Intelligence in Medicine, ***138***, 102476. https://doi.org/10.1016/j.artmed.2022.102476

Zhang, Y., Zhong, P., Jie, D., Wu, J., Zeng, S., Chu, J., Liu, Y., Wu, E. X., & Tang, X. (2021). Brain Tumor Segmentation From Multi-Modal MR Images via Ensembling UNets. ***Frontiers in Radiology***, ***1***, 704888. https://doi.org/10.3389/fradi.2021.704888

Zhang, Z., Yao, L., Wang, B., Jha, D., Durak, G., Keles, E., Medetalibeyoglu, A., & Bagci, U. (2024). DiffBoost: Enhancing Medical Image Segmentation via Text-Guided Diffusion Model. ***IEEE Transactions on Medical Imaging***, 1–1. https://doi.org/10.1109/TMI.2024.3519307

Zhang, Z., Wang, Z., Lin, Z., & Qi, H. (2019). Image Super-Resolution by Neural Texture Transfer, 7974–7983. https://doi.org/10.1109/CVPR.2019.00817

Zhou, C., Ding, C., Wang, X., Lu, Z., & Tao, D. (2020). One-Pass Multi-Task Networks With Cross-Task Guided Attention for Brain Tumor Segmentation. ***IEEE Transactions on Image Processing***, ***29***, 4516–4529. https://doi.org/10.1109/TIP.2020.2973510

Zhou, C., & Paffenroth, R. C. (2017). Anomaly Detection with Robust Deep Autoencoders. ***Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining***, 665–674. https://doi.org/10.1145/3097983.3098052

Zhou, H., Huang, Y., Li, Y., Zhou, Y., & ZhengFellow, Y. (2022). Blind Super-Resolution of 3D MRI via Unsupervised Domain Transformation. ***IEEE journal of biomedical and health informatics***, ***PP***. https://doi.org/10.1109/JBHI.2022.3232511

Zhou, Q., & Zou, H. (2022). A layer-wise fusion network incorporating self-supervised learning for multimodal MR image synthesis. ***Frontiers in Genetics***, ***13***. https://doi.org/10.3389/fgene.2022.937042

Zhou, X., Yamada, K., Kojima, T., Takayama, R., Wang, S., Zhou, X., Hara, T., & Fujita, H. (2018). Performance evaluation of 2D and 3D deep learning approaches for automatic segmentation of multiple organs on CT images.

Medical Imaging 2018: Computer-Aided Diagnosis, ***10575***, 520–525. https://doi.org/10.1117/12.2295178

Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., & Shao, L. (2019). Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images. ***2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 2074–2083. https://doi.org/10.1109/CVPR.2019.00218

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). ***Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks***.

Zhu, Z., He, X., Qi, G., Li, Y., Cong, B., & Liu, Y. (2023). Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. ***Information Fusion***, ***91***, 376–387. https://doi.org/10.1016/j.inffus.2022.10.022

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. ***Proceedings of the IEEE***, ***109***(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

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***, 289–297. https://doi.org/10.1007/978-3-030-32251-9_32

Zong, H., Parada, L. F., & Baker, S. J. (2015). Cell of Origin for Malignant Gliomas and Its Implication in Therapeutic Development. ***Cold Spring Harbor Perspectives in Biology***, ***7***(5), a020610. https://doi.org/10.1101/cshperspect.a020610

Zumel-Marne, A., Kundi, M., Castaño-Vinyals, G., Alguacil, J., Petridou, E. T., Georgakis, M. K., Morales-Suárez-Varela, M., Sadetzki, S., Piro, S., Nagrani, R., Filippini, G., Hutter, H.-P., Dikshit, R., Woehrer, A., Maule, M., Weinmann, T., Krewski, D., 't Mannetje, A., Momoli, F., ... Cardis, E. (2020). Clinical presentation of young people (10-24 years old) with brain tumors: Results

from the international MOBI-Kids study. ***Journal of Neuro-Oncology***, ***147***(2), 427–440. https://doi.org/10.1007/s11060-020-03437-4